

Peer Review File

Manuscript Title: Heat Assisted Detection and Ranging

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

In this paper, the authors propose a novel state-of-the-art HADAR modality for detection and ranging focused on the thermal region of the light spectrum. To address the ghosting effect in thermal radiation, the authors show that TeX decomposition extracts the texture and depth information to estimate temperature and emissivity from the thermal data. Exploiting the fully passive property of thermal radiation from objects, the authors develop a promising estimation theory addressing shot-noise limits to achieve state-of-the-art performance in various ML tasks with physics-driven models. The authors primarily use three major attributes, temperature, emissivity, and texture to describe the thermal images. To demonstrate the efficacy of the proposed HADAR framework, the authors pick major computer vision tasks such as object detection, depth estimation, semantic segmentation, and automated thermography.

The fundamental contribution of this paper is to recover the texture information from thermal radiation by breaking TeX degeneracy and constructing a custom material library of spectral emissivity to train a Neural Network to estimate TeX information for HADAR. In addition, the authors also propose HADAR estimation theory to address limitations of thermal signatures-based ML tasks and demonstrate that physics governed ML decisions are more accurate than data-driven ML models for thermal imagery-related tasks. For ML tasks, the authors use the following extracted parameters: temperature, emissivity, and texture jointly as done equivalently with color, brightness, and saturation for the RGB domain. The reviewer appreciates the author's rigorous work in the mathematical illustration and physical significance of the underlying process discussed in the paper. Some of the promising applications of this framework would be in the robotic vision for self-driving cars at night or to aid search and rescue operations such as firefighters utilizing thermal imagery to interpret their surroundings more effectively.

The reviewer has some concerns about this paper which are discussed below.

1. It would be interesting to see what the neural network learns on training from the material library? Saliency maps visualization would be helpful and some correlation with physical attributes of thermal data. Also, having a physics-informed ML model with a physics-based loss function would be insightful here. Even though the authors call this NN a physically aware machine perception, having the NN train without such a physics model-based loss function doesn't justify the objection. Adding the above-suggested results would make the paper more compelling.

2. The authors mention a very detailed comparison with the state-of-the-art thermal imaging in the paper, The reviewer thinks they are limited and not quantified well for each task. Detailed quantification of the performance of HADAR vs state of art method with metrics such as accuracy, AUC, or mAP for each task would be more credible to readers and justify the efficacy of the HADAR.

3. For the scalable performance of HADAR based on this TeX decomposition, is there any form of incremental learning to address the estimation issue as the material library grows per the demand of the user? Will there be performance degradation as the library grows? How does the efficacy analysis using the Cramer Rao Bound change in such a scenario?

4. The authors have demonstrated the Estimation theory under theoretical and simulation grounds which is a good starting point. However, it is equally important to demonstrate the efficacy of this proposed method on a real-world dataset. How do the authors do HADAR ranging and stereo vision? A more in-depth description of this process would aid clarity of the paper and understanding of the reader. Also, what is meant by ~ 100 x accuracy in ranging? Does this apply to all scenarios rather than specific cases?

5. As per the authors, the HADAR framework requires a hyperspectral cube to input to the NN. Despite such promising performance of HADAR, its applicability on real-world datasets might be limited as most of the real-world thermal datasets are only available based on temperatures. Do authors intend to extend the HADAR estimation to extrapolate the rest of the information based on this limitation in information availability for more applicability?

6. The NN devised for TeX decomposition is a flattened fully connected layer/1D CNN, which results in the loss of spatial-spectral information during the inferences. How much of that is evident in current results? Quantification of results is missing and needs to be added to make the argument for HADAR application more compelling. Will a 3D CNN based on spatial-spectral info be more useful for better performances?

7. It is not clear if the authors can identify pixels corresponding to different material from 1D CNN. Is the material identification done at a global image-level or pixel-by-pixel basis? Would 1D CNN suffice for the task if it is done pixel by a pixel basis? Something like UNET based models would be apt for such tasks.

8. In the supplementary material for HADAR estimation theory, in the last sentence of the section, Theory of Texture, the authors mention that HADAR texture is equivalent to grayscale imaging in daylight, which the reviewer thinks is False as grayscale imaging possesses greater qualitative and quantitative texture information than that of HADAR as inferred by the results currently presented in the paper. Improving results to substantiate this claim or else removing it is suggested.

9. To prove the efficacy of the HADAR, the authors estimate material characteristics with NN but use a direct inverse function to estimate T and X. How effective would the application of such inverse function be for real-life applications in comparison to the simulations? Such functions sound promising for theoretical grounds and simulation but involve many constraints and noise factors for real-world application.

10. For the thermography and semantics experiments, the authors utilize data duplication instead of augmentation to construct a larger dataset. Such methods are not encouraged for DL applications as they might bias the NN towards a certain class. Rather, the application of augmentation techniques is preferred.

11. Regarding the applicability of the proposed HADAR setup for application, for dynamic environments such as self-driving vehicles, it would be expensive to have a multispectral acquisition device that can simultaneously acquire the data in the multi-spectrum. Even for just proof of concept in this HADAR framework, the authors change filters via a wheel to retrieve the spectral resolution which may need switching between filters for data acquisition and camera stabilization with such abruptly varying acquisition windows. Such a technique might be problematic in dynamic environments yielding a lot of background noise in TeX decomposition. Suggest adding a paragraph that addresses such current constraints on the real-world application of the method.

The authors present a promising and potentially groundbreaking new methodology in their presentation of HADAR. They present significant improvements in performance over other modalities in low light conditions. However, such a modality is not without significant challenges still to be addressed before it can be recognized as a “next step” in computer vision applications. From on-fly calibration to the design of acquisition devices pose hardware-level challenges. Besides, the interface of such modules with edge computing devices for real-world applications would be a challenge where the TeX decomposition framework and Task-based frameworks can be easily be deployed in such devices. Another challenge is coming up with a robust library to train the framework as the material properties also change with the environment leading to change in each TeX parameter. Also, the authors don't present the acquisition between cold and hot conditions. How that changes HADAR performance? Also, for cost-effectiveness, most of the available thermal cameras used in day-to-day life on consumer products are low priced. The applicability of such a multispectral camera for the HADAR application may not be cost-effective and affordable to low-end consumer products and even

academic research. The authors make significant projections about the future of this work and have done excellent work in their theory. However, they have neglected to address any of the real and significant challenges that remain before the implementation of such work can take place in a real-world application. Suggest the addition of a section that highlights remaining constraints on the work before it can be presented as a real-world solution. This could take the form of a paragraph in the Discussion or Outlook. The reviewer believes that HADAR brings a lot of potential to the world but as it currently stands, bears significant challenges that need to be carefully addressed before it can replace or substitute the existing modalities in decision making.

Referee #2 (Remarks to the Author):

A. Summary of key results.

The state-of-the-art machine perception utilizing active sonar, radar and LiDAR to enhance camera vision is not viable as the number of intelligent agents scales up. Exploiting omnipresent heat signals could be a new frontier for scalable perception. However, objects and their environment constantly emit and scatter thermal radiation leading to textureless images famously known as the 'ghosting effect'. In this work, the authors proposed a method called HADAR to overcome this ghosting effect by decomposing the heat signal into temperature, emissivity and texture (TeX decomposition). They have developed the HADAR estimation theory and address its shot-noise limits depicting information-theoretical bounds to HADAR-based AI performance. In addition, they have also developed HADAR ranging (depth estimation) that shows an accuracy improvement up to two orders of magnitude compared with existing thermal ranging. They have performed physics-driven semantic segmentation to achieve improved performance against AI-enhanced thermal sensing.

B. Originality and Significance

This article focuses on the separability of temperature and emissivity from a thermal signal and the use of emissivity profiles for detection and ranging. The separability is discussed earlier in literature and used for various computation assisted tasks. Cramer-Rao bound on the distance for identifiability based on intrinsic properties of a material ascertains quantification of the utility of thermal imaging. Whereas an error bound given on the ranging accuracy with a limitation on photons counts further determines the accuracy of perception.

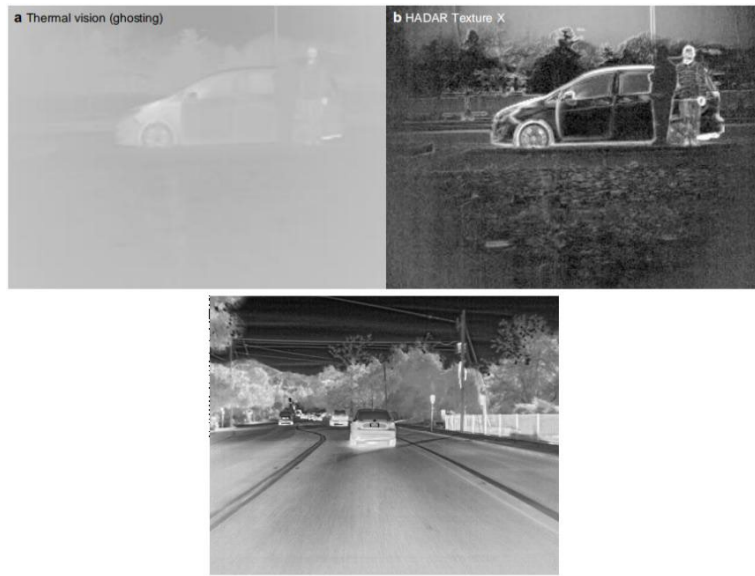
The third component of significance is texture, which is computed from emissivity. Texture in visible light imaging qualifies the identification. TeX decomposition in thermal signal allows fine distinguishable parameters for image processing.

C. Data and Methodology

There are few comments about Data and Methodology as follows:

1. A schematic diagram of the hardware setup would benefit the readers. Though an image of the hardware setup is provided, it seems insufficient for a scholarly article.
2. Authors have derived the Cramer-Rao bound for HADAR estimation and the machine learning method but have not provided proof for the Cramer-Rao bound.

3. Authors have selected a single environment for experimentation setup with a few characteristically distinguishable objects (person and cardboard Einstein model). It is evident from the computer vision literature that state-of-the-art artificial intelligence algorithms perform comparatively better in complex scenes. So the efficacy of the proposed method is hard to determine.
4. For the development of HADAR, they have used FLIR A325sc, which is outdated. In addition, they have made the comparison with 16 channel lidar Velodyne puck, which gives sparse data. A comment about the functionality and cost comparison would help readers appreciate the proposed technique's significance.
5. Authors state, "HADAR is distinct from hyperspectral imaging where material difference is determined by the Euclidean distance between their reflectance spectra [32]. In stark contrast, HADAR identifiability is determined by multi-parameter estimation of temperature, emissivity and texture", but identifiability is estimated using a CNN with an input of proton profiles only (as given in the Methodology section). A precise statement would help readers to understand the implementation details.
6. Authors claim, "The minimum photon number for given semantic distance or vice versa, the minimum semantic distance for given photon number sets fundamental limits to object identification beyond training volume, providing a theoretical foundation for designing public policies." However, ML algorithms generally perform well with missing or scarce information; placing a bound on input data is yet an open challenge. A comment about the significance of bounds would help the cross-discipline reader base.
7. Authors said that "Thermal imaging loses textures due to TeX degeneracy (Fig. 4a) and leads to inaccurate ranging". Thermal cameras, for instance, FLIR BlackFly (BFS-U3-51S5C-C), produce impressive images with texture, as shown in figure-1. Moreover, in this work, emissivity is used for range computation instead of texture.
8. An example where thermal equilibrium can cause singularity would help appreciate the utility of identifiability and ranging.
9. A comment about the change in-bounds considering the non-stationary objects in a scene would help assess TeX utility in interpreting sequential information.



From FLIR Dataset

Figure-
1

D. Appropriate use of statistics and treatment of uncertainties

1. They have claimed that they have achieved 100 x accuracy in HADAR ranging, which is physics-based semantic segmentation between a person and a metallic body. In regard to computer vision literature, AI-based semantic segmentation results are already established. They have not made a comparative analysis between their proposed method and state-of-the-art AI-based semantic segmentation. Second, they have done the semantic segmentation using emissivity, and if the two subjects have the same emissivity, then their method fails. Below are some references for the AI- based semantic segmentation on thermal images
 - Li, Chenglong, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation." *IEEE Transactions on Neural Networks and Learning Systems* (2020).
 - He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969. 2017.
 - Treible, Wayne, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu.

- "Cats: A color and thermal stereo benchmark." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2961-2969. 2017.
2. In the HADAR ranging, the authors claim that they have used several AI algorithms for instance, DeepPruner, PSMNet, but they have not provided any training details for these algorithms. Similarly, details of data collection and experimental results for these algorithms are not found in this manuscript. Qualitative and quantitative results would help readers and reviewers to make a fair comparison.
 3. *"We develop HADAR estimation theory to address fundamental limits of object identification from its thermal infrared signature. We believe this will be crucial in guiding public policy for the industrial revolution where decision accuracy of machine perception can be bounded by physical laws as opposed to training data volume"*. It would help readers and reviews to appreciate the claims if an example or two are provided where the proposed technique helps public policy.

E. Suggested improvement

In addition to the above comments, some additional comments are as follows that require more explanation.

1. *"where multiple attributes are desired either for safety guarantees or scientific purpose"*. Some examples with reference are required for this claim.
2. *"However, large scale temperature screening with existing noncontact infrared thermometer or infrared thermography is ineffective due to lack of adaptivity to emissivity (complexion/makeup), distance, age, gender, and circadian variations [36–38]."* A brief comment on the utility of TeX, in this case, will help the reader to appreciate
3. *"Cramér-Rao bound is therefore promising for the smart healthcare industry including early reliable skin cancer detection."* Reference and a brief comment will help users to understand the relation
4. *"Our results call for heat exploitation in the quantum regime where single photon detectors are being developed in the thermal infrared"*. It is not mentioned in the whole script except in the introduction, a bit of explicit comment may help the readers.
5. *"However, large scale temperature screening with existing non contact infrared thermometer or infrared thermography is ineffective due to lack of adaptivity to emissivity (complexion/makeup), distance, age, gender, and circadian variations' "*. Reference and a brief comment will help users to understand the relation.
6. They mention a *"phantom breaking phenomenon"* as a disadvantage of thermal imaging, but do not explain if the proposed technique addresses it.

F. References

1. *"The emerging Industry 4.0 of smart technologies [18] calls for a future with scalable human-robot social interactions since it is expected that one in ten vehicles will be automated by 2030 and 100 million robot helpers will be serving people."* The reference paper has no such claim.
2. *"Scalable perception"*. No explanation is given for the scalable perception
3. They claimed that this method is novel but the following are the works that have done temperature emissivity separation.
 - Jie Cheng, Qing Xiao, Xiaowen Li, Qinhuo Liu, Yongming Du, Aixiu Nie, "Multi-layer perceptron neural network based algorithm for simultaneous retrieving temperature and emissivity from hyperspectral FTIR dataset", Geoscience and Remote Sensing Symposium 2007. IGARSS 2007. IEEE International, pp. 4383-4385, 2007.

- Xinghong Wang, Xiaoying OuYang, Bohui Tang, Zhao-Liang Li, Renhua Zhang, "A New Method for Temperature/Emissivity Separation from Hyperspectral Thermal Infrared Data", Geoscience and Remote Sensing Symposium 2008. IGARSS 2008. IEEE International, vol. 3, pp. III - 286-III - 289, 2008.
- Hang Yang, Lifu Zhang, Junyong Fang, Xia Zhang, Qingxi Tong, "Algorithm research of building materials emissivity extracting", Geoscience and Remote Sensing Symposium (IGARSS) 2010 IEEE International, pp. 3350-3353, 2010.
- Hang Yang, Lifu Zhang, Li Liu, Qingxi Tong, "Temperature and emissivity separation from TASI data based on wavebands selection", Geoscience and Remote Sensing Symposium (IGARSS) 2011 IEEE International, pp. 1850- 1853, 2011.
- Ning Wang, Yonggang Qian, Hua Wu, Lingling Ma, Zhao-Liang Li, Lingli Tang, "Performances of temperature and emissivity separation methods for hyperspectral thermal data affected by the changes of spectral properties of sensor", Geoscience and Remote Sensing Symposium (IGARSS) 2013 IEEE International, pp. 2152-2155, 2013.
- Schmugge, Thomas, Andrew French, Jerry C. Ritchie, Albert Rango, and Henk Pelgrum. "Temperature and emissivity separation from multispectral thermal infrared observations." *Remote Sensing of Environment* 79, no. 2-3 (2002): 189-198.
- V. Payan Corresponding author & A. Royer (2004) Analysis of Temperature Emissivity Separation (TES) algorithm applicability and sensitivity, *International Journal of Remote Sensing*, 25:1, 15-37, DOI: 10.1080/0143116031000115274

G. Clarity

1. Language use in writing is a bit extreme (e.g. "*that can disrupt AI industry*", "*TeX degeneracy*")
2. The authors refer to the supplementary information frequently but do not mention the section which becomes bothersome for the reader.

H. Decision

1. The proposed technique is a fascinating idea and can benefit the AI researchers with another reliable sensor.
2. However, the current article is not ready for publication in its current form. It is suggested to consider the recommendations and resubmit.

Author Rebuttals to Initial Comments:

Cover letter to Reviewer 1

We would like to thank the reviewer for the encouraging response and valuable comments. Here, we list all the major revisions, and we will provide individual replies to each comment from the next page onwards.

Reviewers' main concerns and corresponding major revisions include that

Problem (1):

Details of our machine learning, such as, network architecture, data preparation, training analysis, quantitative comparisons of HADAR performances (semantic segmentation, detection, and ranging) with the state-of-the-art, are not provided.

Revision (1):

We have expanded the Supplementary Information and Methods to provide more details about our HADAR theory, machine learning, and experiments. Especially, we have explained the architecture of our TeX-Net for TeX decomposition, Saliency maps in material classification, and physics-based loss. We have also made quantitative and qualitative comparisons of machine learning performances based on our TeX vision against the state-of-the-art thermal vision. In particular, our results focus on semantic segmentation, people detection, and ranging.

Problem (2):

The previous version only demonstrated HADAR efficacy for a few simple scenes. HADAR efficacy for real-world level complicated scenes is not verified.

Revision (2):

We have built and released the 1st HADAR database with complicated scenes and clearly shown HADAR efficacy.

Problem (3):

HADAR efficacy on hot and cold weather conditions are not compared.

Revision (3):

We have also added one more experiment in summer daylight to compare HADAR performances on cold winter and hot summer conditions.

Problem (4):

Textures recovered in HADAR are not quantified and compared to the state-of-the-art approaches.

Revision (4):

We have quantified textures and have made comparison with state-of-the-art approaches to show the advantage of HADAR in recovering textures.

We have also made revisions according to all other comments. Now, we will address each comment sequentially in the following. Notations used in this response include C: Comment, R: Reply, *Italic*: revisions, underline: emphasize.

Reviewer 1	
C0	<p>In this paper, the authors propose a novel state-of-the-art HADAR modality for detection and ranging focused on the thermal region of the light spectrum. To address the ghosting effect in thermal radiation, the authors show that TeX decomposition extracts the texture and depth information to estimate temperature and emissivity from the thermal data. Exploiting the fully passive property of thermal radiation from objects, the authors develop a promising estimation theory addressing shot-noise limits to achieve state-of-the-art performance in various ML tasks with physics-driven models. The authors primarily use three major attributes, temperature, emissivity, and texture to describe the thermal images. To demonstrate the efficacy of the proposed HADAR framework, the authors pick major computer vision tasks such as object detection, depth estimation, semantic segmentation, and automated thermography.</p> <p>The fundamental contribution of this paper is to recover the texture information from thermal radiation by breaking TeX degeneracy and constructing a custom material library of spectral emissivity to train a Neural Network to estimate TeX information for HADAR. In addition, the authors also propose HADAR estimation theory to address limitations of thermal signatures-based ML tasks and demonstrate that physics governed ML decisions are more accurate than data-driven ML models for thermal imagery-related tasks. For ML tasks, the authors use the following extracted parameters: temperature, emissivity, and texture jointly as done equivalently with color, brightness, and saturation for the RGB domain. The reviewer appreciates the author's rigorous work in the mathematical illustration and physical significance of the underlying process discussed in the paper. Some of the promising applications of this framework would be in the robotic vision for self-driving cars at night or to aid search and rescue operations such as firefighters utilizing thermal imagery to interpret their surroundings more effectively.</p> <p>The reviewer has some concerns about this paper which are discussed below.</p>
R0	<p>We would like to thank the reviewer for the encouraging response and valuable comments. We have addressed each comment individually below and made major revisions to improve the quality of this manuscript.</p>
C1	<p>It would be interesting to see what the neural network learns on training from the material library? Saliency maps visualization would be helpful and some correlation with physical attributes of thermal data. Also, having a physics-informed ML model with a physics-based loss function would be insightful here. Even though the authors call this NN a physically aware machine perception, having the NN train without such a physics model-based loss function doesn't justify the objection. Adding the above-suggested results would make the paper more compelling.</p>
R1	<p>We thank the reviewer for pointing this out. We agree with the reviewer that a physics-based loss function is essential to call HADAR TeX-Net physically aware.</p> <ol style="list-style-type: none"> 1. In this new version of the manuscript, <u>we've added the architecture of our TeX-Net for TeX decomposition, as well as the physics-based loss function (Extended Data Fig.1), see below.</u>

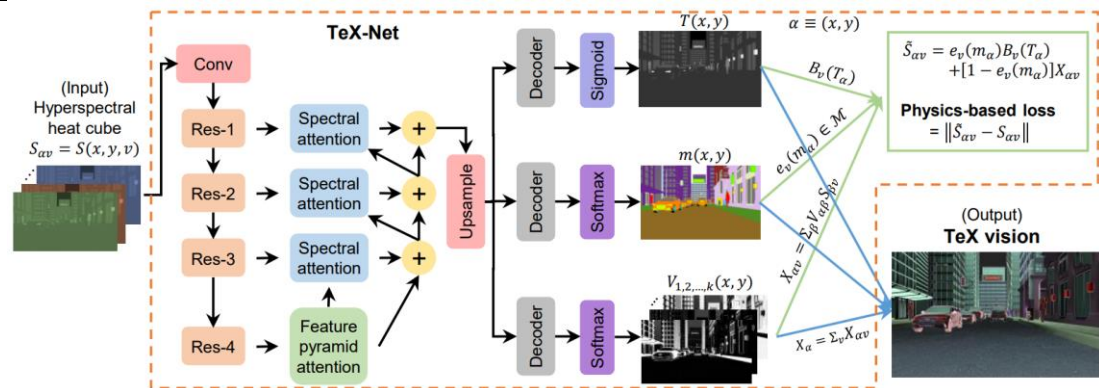


Fig.1 Architecture of TeX-Net for TeX decomposition. The input to TeX-Net is the hyperspectral heat cube. The output is the TeX vision. Physics-based loss function to train TeX-Net is defined on the reconstructed heat cube, which is based on physics models of blackbody radiation, the material library, and the mathematical structure of texture X. This figure is given in Extended Data Fig.1.

For TeX decomposition, we need to extract information from both spatial patterns and spectral thermal signatures. This motivates us to use spectral and pyramid attention layers in the UNet model (PAN [1]). Even though TeX-Net outputs TeX vision, one cannot directly use ground truth TeX to train the network. The mathematical structure of X has to be specified to ensure the uniqueness of inverse mapping and overcome TeX degeneracy. Hence, our key insight is to learn thermal lighting factors V instead of texture X. Texture X is constructed with thermal lighting factor V indirectly.

Supervised Learning: In supervised learning, TeX-Net is trained with ground truth temperature (T), material index (m), and thermal lighting factor (V). The ground truth T, m, and V are obtained with least-squares estimation based on the material library. The loss function is a combination of individual losses with regularization hyper-parameters.

Unsupervised Learning: In un-supervised learning, no ground truth data is required. The material library is built into the network. The physics-based loss function defined on the re-constructed heat cube is based on physics models of the heat signal. The network is trained with material library \mathcal{M} , Planck's law $B(T)$, and the mathematical structure of texture (X). In practice, we use a hybrid loss function with T, e, V contributions in addition to the physics-based loss.

2. In this paper, we have built the city block dataset to train TeX-Net. The HADAR-CityBlock dataset is the first LWIR (long-wave infrared) stereo-hyperspectral dataset in the world with ground truth depth and ground truth TeX vision. We have made the dataset available at https://drive.google.com/drive/folders/1da2Uh5t_QOy-MrWxhkJJw3MueNxsuVtn?usp=sharing. We will host it on Github for the research community once the paper is published.
3. We have trained the TeX-Net with supervised, un-supervised, and hybrid learning. Results of supervised learning is given in Fig.S12 in Sec.SIIIA of the Supple. Info. The Saliency map for material e(m) is given in Fig.S13 in in Sec.SIIIA, as cited below.

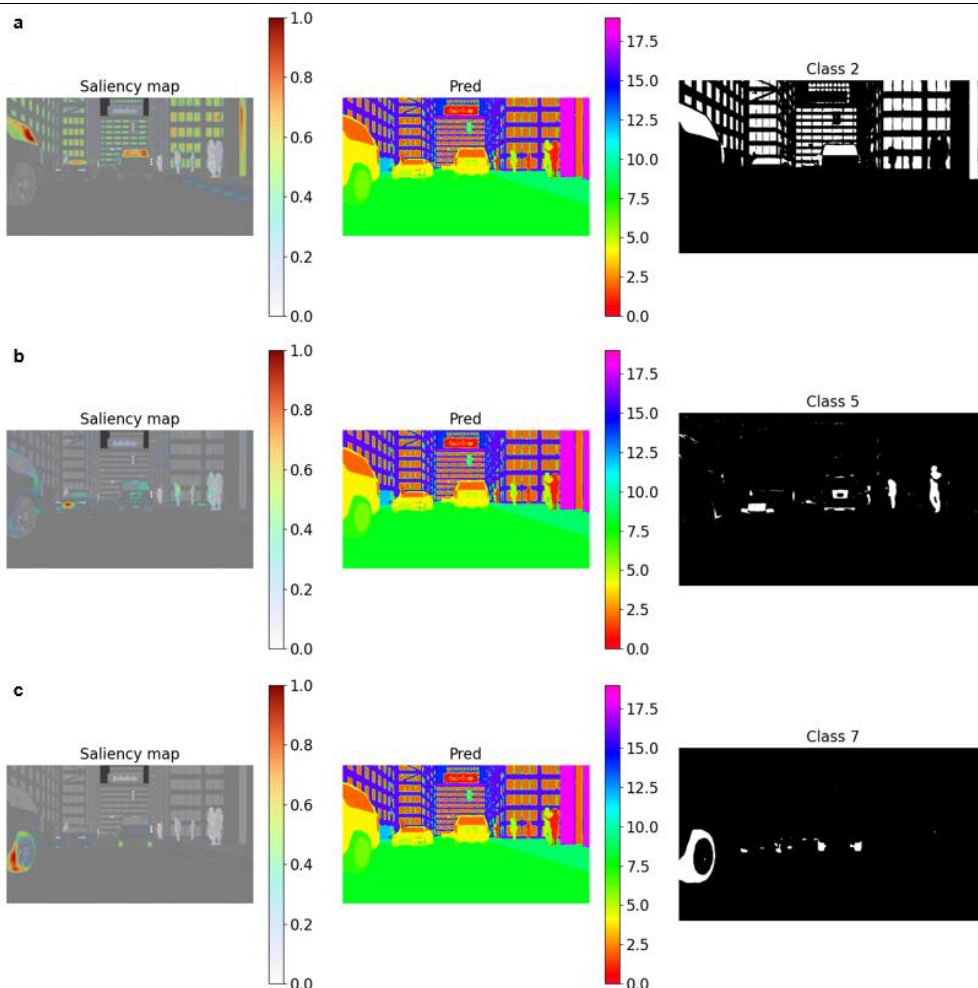


Fig.2 Saliency map of TeX-Net in supervised learning. The active region in Saliency maps is localized and highly correlated with the corresponding material region (last column), indicating that TeX-Net has properly learnt spatial and spectral features for material classification. 3 samples out of 20 materials are shown. a, Saliency map for class 2, window glass. b, Saliency map for class 5, aluminum. c, Saliency map for class 7, tire. Pred: material index prediction of TeX-Net. This figure is the Fig.S13 in Supple. Info.

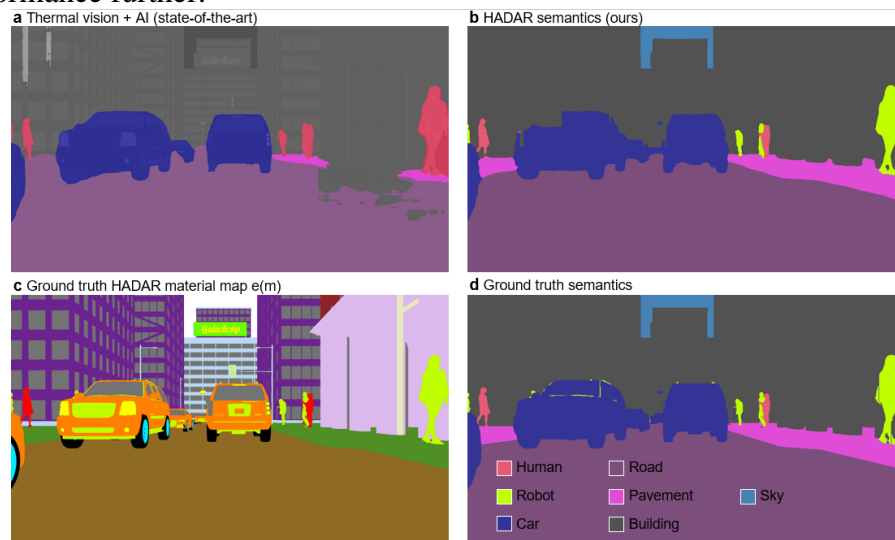
Reference(s):

[1] Li, Hanchao, et al. "Pyramid attention network for semantic segmentation." arXiv preprint arXiv:1805.10180 (2018).

C2	The authors mention a very detailed comparison with the state-of-the-art thermal imaging in the paper, The reviewer thinks they are limited and not quantified well for each task. Detailed quantification of the performance of HADAR vs state of art method with metrics such as accuracy, AUC, or mAP for each task would be more credible to readers and justify the efficacy of the HADAR.
R2	We thank the reviewer for asking about the quantitative comparison of our HADAR performance with the state-of-the-art AI-enhanced thermal sensing. We agree with the reviewer that

quantification of performance enhancement is necessary to justify HADAR efficacy. In this new version, we have added quantitative and qualitative comparisons, regarding people detection, semantic segmentation, and ranging, to show the advantages of HADAR TeX vision over traditional thermal vision.

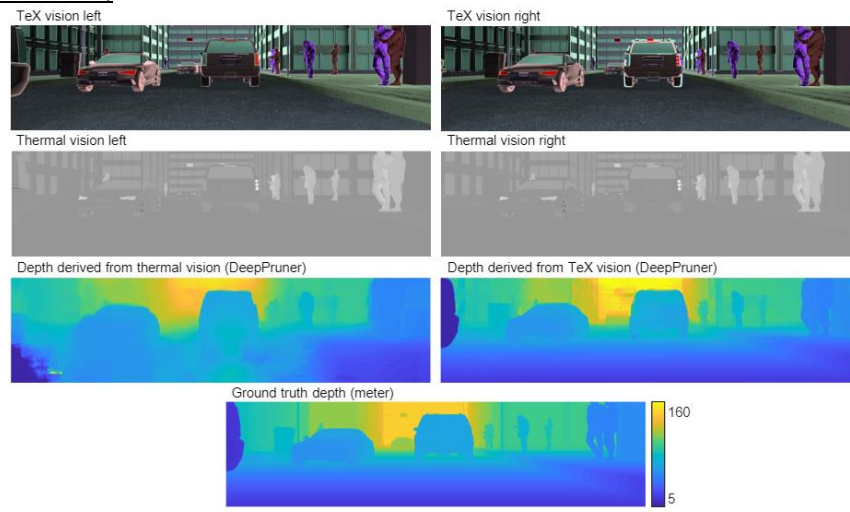
1. Before quantifying the performance, we would like to explain the ground truth we used. We have built the 1st LWIR stereo-hyperspectral dataset in the world, HADAR-CityBlock, as the platform to test HADAR efficacy. The ground truth depth and ground truth TeX vision of the city block scene are synthesized by Monte Carlo path tracing, through exploiting Planck’s law and Kirchhoff’s law in Blender Cycles. Ground truth semantic segmentation is generated by transforming the material map in TeX vision with a customized algorithm, see algorithm 4 of the Supple. Info. Our demonstration of physics-driven detection focuses on distinguishing similar visual appearances (e.g., human vs. robot) where conventional techniques exploiting visual appearance alone would fail. In two separate scenes, we will show HADAR advantage of using material signature for people detection. We do not provide detailed statistics of people detection since the conventional vision -driven techniques fail.
2. To quantify semantic segmentation performance, the metric of mIoU (mean intersection over union) is used instead of mAP, according to Ref. [1] on DANet. We used the DANet (pre-trained on the Cityscapes dataset) on thermal vision as the state-of-the-art baseline, and we used the output of TeX-Net plus our algorithm 4 in Supple. Info. to get HADAR semantic segmentation. The quantitative comparison is given in Extended Data Fig.8, as cited below. We note that we used AI-enhancement only in thermal vision but not in HADAR. HADAR semantics outperforming thermal semantics clearly show that the advantage of HADAR semantics comes from TeX vision but not the AI algorithm. In the future, dedicated AI algorithms can be developed using our HADAR database to improve the performance further.



mIoU %	Human	Robot	Car	Road	Pavement	Building	Sky
Thermal vision + AI	35	0	83	85	39	92	0
TeX vision + non-AI	82	79	91	99	96	98	95

Fig.3 HADAR TeX-physics-driven semantic segmentation beats state-of-the-art thermal-vision-driven semantic segmentation (thermal vision + AI). a, Thermal semantic segmentation with DANet (pre-trained on the Cityscapes dataset). b, HADAR semantic segmentation transformed from the material map in estimated TeX vision. c, Ground truth material map in the ground truth TeX vision. d, Semantic segmentation transformed from (c) to approximate the ground truth segmentation, see Sec.SIIIE of the Supple. Info. for more details of the non-machine-learning transformation. Since AI enhancement is only used in thermal semantics, the advantage of HADAR semantics is clearly from TeX vision with physical attributes. Statistics in the table is analysed for 10 frames (8:10:98) of the left camera in the city block dataset 1. mIoU: Pixel-wise mean intersection over union.

- To quantify stereo matching performance, we used the metrics of mean disparity error (mean absolute per-pixel disparity error with respect to the ground truth) and Accuracy (fraction of pixels for which the estimated disparity is within tau pixels of the ground truth values). We used DeepPruner (pre-trained on the KITTI dataset) as the state-of-the-art AI algorithm. The quantitative and qualitative comparisons are given in Extended Data Figs. 6 and 7, as cited below.



		Density	Mean error (px)	Accuracy (%)			
				$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 10$
Thermal vision + AI	Street	0.5	99.74	4.45	12.65	17.01	22.29
	Entire image	1	55.32	31.54	45.99	50.36	54.98
TeX vision + AI	Street	0.5	1.49	24.74	96.94	98.77	99.66
	Entire image	1	2.09	42.20	93.59	96.64	98.45

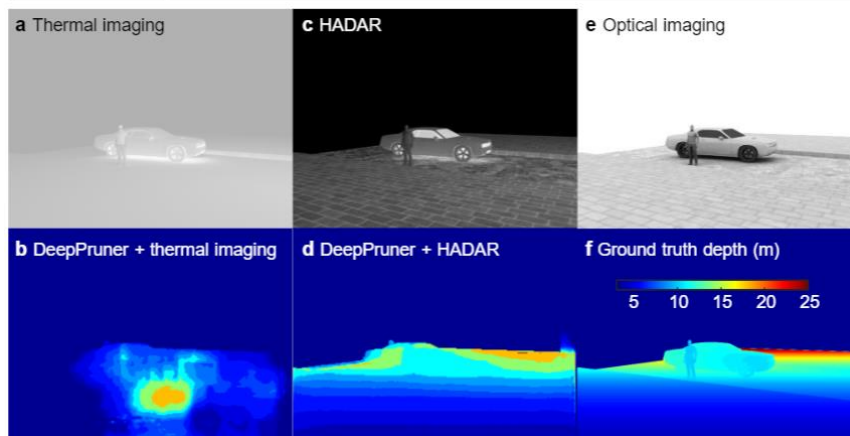


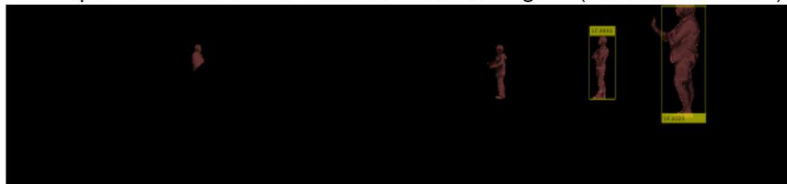
Fig.4 Quantitative and qualitative comparisons of ranging performances based on HADAR and traditional thermal vision. Top figure is the Extended Data Fig.6. 67 folds mean disparity error suppression by HADAR has been observed on the street part, compared with AI-enhanced thermal ranging, and 26 folds mean disparity error suppression has been observed for the entire image. Bottom figure is the Extended Data Fig.7. The fundamental reason for this improved performance in HADAR is due to breaking of TeX degeneracy and overcoming the ghosting effect.

- For object detection, in addition to the previous people detection (current Extended Data Fig.5), we further provided one more figure to demonstrate how HADAR detection can distinguish human vs. robot utilizing material signatures (Fig.S17 of the Supple. Info), which is **impossible** for traditional thermal vision. We note that in HADAR **TeX vision, the scene is captured with physical attributes being represented by hue (material index), saturation (temperature) and value (texture)**. This novel representation has information content which is not present in the output of optical cameras (RGB vision), conventional IR thermal cameras (panchromatic thermal vision), or LiDAR (point cloud). Basically, we extract the material region corresponding to the desired target and only perform detection over the selected region, as shown below. For human detection, we extract the region of material ‘human’ from the material map, and the detection result gives bounding boxes of humans. For robot detection, we extract the region of material ‘aluminum’. Even though car logo and other components of the car are also selected, people detection finds the correct spatial patterns and succeeds in robot detection. In the above case of people detection, we used HOG+SVM pre-trained on the INRIA Person dataset available in Matlab.

a TeX vision



b People detection in the ‘human’-material region (human detection)



c People detection in the ‘aluminum’-material region (robot detection)



Fig.5 Demonstration of physics-driven object detection. HADAR can perform object detection over particular material regions, and hence HADAR can distinguish similar geometries with material signatures. This figure is given as Fig.S17 of the Supple. Info

	<p>5. The previous Extended Data Tab. I was intended to give a qualitative summary of the potentials for HADAR (TeX vision + AI) and traditional thermal sensing (thermal vision + AI). To make our argument more precise, we have revised the caption ‘HADAR outperforms AI-enhanced thermal sensing in detection and ranging. HADAR provides intrinsic physical attributes and enhanced textures enabling comprehensive understanding of the scene beyond AI-enhanced conventional thermal imaging...’</p> <p>6. In this new version, we have also quantified textures and made a quantitative comparison to state-of-the-art approaches. See Sec.SIID of the Supple. Info. for details of texture quantification. See Extended Data Figs. 2, 3, 10, and Fig.S11 of the Supple. Info. for quantitative and qualitative comparisons of textures.</p> <p>Quantifying HADAR performances for all other tasks in computer vision, such as, optical/scene flow, instance segmentation, pose estimation, etc. will be the subject of extensive future studies.</p> <p>Reference(s): [1] Fu, Jun, et al. "Dual attention network for scene segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.</p>
C3	<p>For the scalable performance of HADAR based on this TeX decomposition, is there any form of incremental learning to address the estimation issue as the material library grows per the demand of the user? Will there be performance degradation as the library grows? How does the efficacy analysis using the Cramer Rao Bound change in such a scenario?</p>
R3	<p>We thank the reviewer for asking about incremental learning and the performance change as the library grows. This is a very good point in applications of HADAR. We shall answer this question by analyzing changes in Cramer-Rao bound first and then discussing our observations in learning and performance changes.</p> <p>1. The HADAR identifiability (Cramer-Rao bound) of a target material in a multi-material library is given by Eq.2 in the main text</p> $I = \log_2 \left[1 + \operatorname{erf} \left[\sqrt{\frac{N d_0^2}{2(1 + \gamma)}} \right] \right], \quad (2)$ <p>with the semantic distance d_0 replaced by the minimum semantic distance of the target material with other materials, see Algorithm 3 of the Supple. Info (briefly cited below).</p> <p>5 Search the minimum semantic distance among all $\{m, m_\alpha\}$ pairs,</p> $d_0^{\min} = \min_m \{d_0(m, m_\alpha)\}, \quad (S33)$ <p>and get the corresponding minimum statistical distance d^{\min} ;</p> <p>The underlying physics of using minimum semantic or statistical distance is that, whether material A can be identified or not depends on material B with the most spectral similarity with material A. The other materials in the library do not directly affect the identifiability since the spectral features are discernible with sufficiently high spectral resolution detectors. HADAR identifiability for multi-material library is demonstrated in Extended Data Fig.4, as cited below.</p>

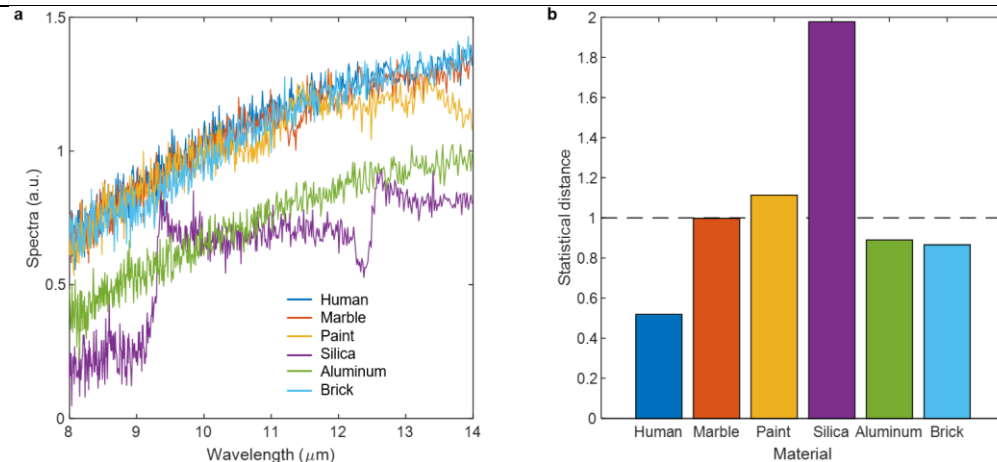


Fig.6 HADAR estimation theory for multi-material library. This figure intuitively shows that HADAR identifiability based on semantic/statistical distance is an effective figure of merit to describe identifiability. a, Incident spectra of 5 materials generated by Monte Carlo simulations. b, Minimum statistical distance of each material. Spectra of silica and paint have non-trivial features that are distinct with other materials in the library. Statistical distance larger than 1 (dashed line) consistently indicates that silica and paint are identifiable. Note that aluminum is similar to human skin under TeX degeneracy and non-identifiable, even though with the same temperature its spectrum is much weaker than human skin.

To see the change of HADAR identifiability with increasing material library, we assume in the beginning we have 2 materials in the library $\mathcal{M}' = \{e_m | m = 1, 2\}$, $m=1$ corresponding to Human and $m=2$ corresponding to Silica, as shown in the above figure. Since Silica has spectral features and is very different with the Human emissivity spectrum, the statistical distance between Human and Silica is around $d(1,2) = 2$, above the distinguishable criterion (dashed line), and hence Human and Silica are both identifiable. This can also be visually seen in Fig.6a.

Now, we introduce a third material into the library, $m=3$ corresponding to Brick. Brick has a spectrum similar with Human but is different from Silica. The statistical distance between Brick and Human is around $d(1,3) = 0.5$ and the statistical distance between Brick and Silica is around $d(2,3) = 2$. The minimum statistical distance of Human from two other materials in the library becomes $d_1 = \min[d(1,2), d(1,3)] = 0.5$, and the minimum statistical distance of Silica from two other materials in the library is $d_2 = \min[d(1,2), d(2,3)] = 2$. This implies that after introducing Brick into the material library, Human becomes non-identifiable, and Silica is still identifiable. After introducing Marble, Paint, and Aluminum into the library, Silica is still identifiable. Overall, the minimum semantic and/or statistical distance of each material in the library will decrease when new materials are introduced into the library. The influence on each material is different, as discussed in the above paragraph. Semantic and statistical distance will increase with better spectral resolution (more spectral bands) in the camera.

The above analysis implies the following scaling law.

- **For fixed spectral resolution, the more materials in a library, the more difficult it is to distinguish each of them. For a fixed number of materials in the library, higher the spectral resolution, the easier it is to distinguish each of them. And to distinguish a large number of materials in a library, higher spectral resolution and low noise sensors are required.** This paragraph has been added in Sec.SIIB of the Supple. Info. to help readers understand the bound changes with increasing materials.

2. We shed light on the performance changes of growing library size by analyzing its role on TeX-Net training loss using our HADAR-Cityblock dataset. In the Cityblock dataset, there are 20 materials in the material library, but we train the network with a subset of the library, i.e., with only a few materials in the library. This scenario is closely related to real-world applications where, in the real-world scene, there might be over hundreds of materials, but we've only calibrated a few of them. These few selected materials amount to be material classes into which we can approximate the scene. Explicitly, we trained TeX-Net with only 3 materials in the library, distinct glass and brass, and all others were approximated as a blackbody. This approximation will surely bias the temperature and texture predictions, but as the number of materials in the library grows, the overall physics-based loss will decrease. We repeat the above training for different number of materials in the library, and indeed, we observe loss decrease as cited below. In contrast to point 1, here, number of materials in the scene is fixed, and more materials in the library leads to better approximation of the scene.

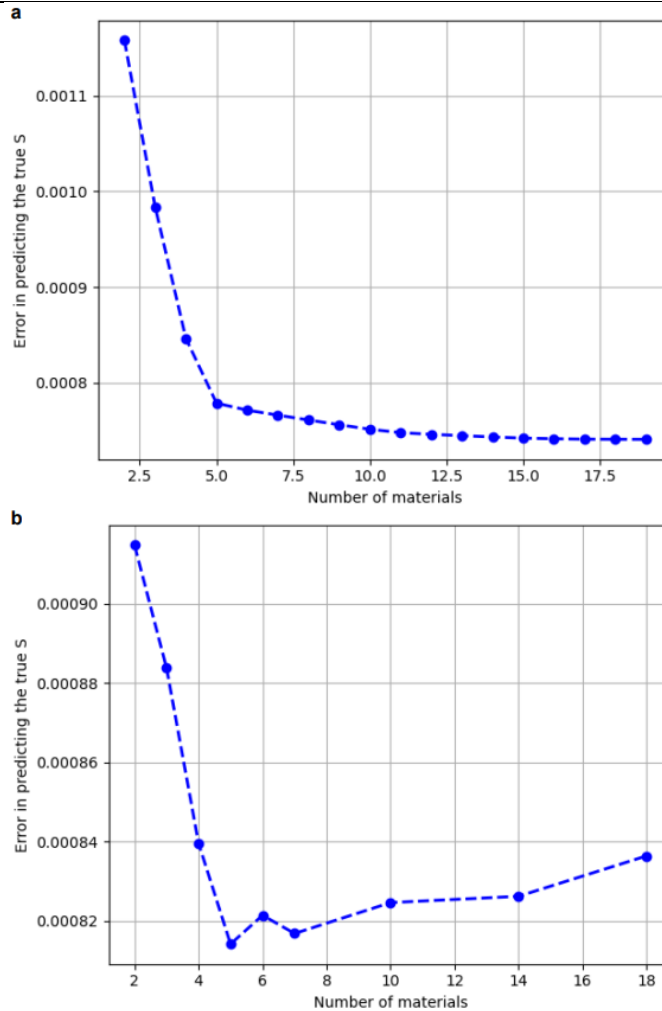


Fig.7 Physics-based loss decreases as the number of materials in the library increases. a, materials are added into the library with a greedy approach, and pixels are classified into those material classes based on visual similarity. Temperature and thermal lighting factors are solved out accordingly. b, Pixels are classified into material classes with neural network (TeX-Net). TeX-Net finds more accurate TeX decomposition, and again, we can see that with more materials in the library the physics-based loss is lower. The error in (b) after 5 materials is noise. This figure is the Fig.S14 of the Supple. Info.

3. We note that the number of epochs needed to train the network is not significantly slower with increasing materials in the library. TeX-Net was trained independently for different number of materials in the library. As incremental learning itself is an open question, we plan to explore it in a detailed and dedicated study in the future.

C4.1 The authors have demonstrated the Estimation theory under theoretical and simulation grounds which is a good starting point. However, it is equally important to demonstrate the efficacy of this proposed method on a real-world dataset.

R4.1 We agree with the reviewer that it is necessary to demonstrate HADAR efficacy with real-world level complicated scenes and compare with state-of-the-art AI-enhanced thermal vision.

However, we note that there is no experimental hyperspectral LWIR database with ground truth depth available in the literature. To further demonstrate HADAR efficacy in this new version, we have: (1) built and released **the 1st synthesized HADAR database with real-world level complicated scenes** (https://drive.google.com/drive/folders/1da2Uh5t_QOy-MrWxhkJJw3MueNxsuVtn?usp=sharing; we will host it on Github for the scientific community once the paper is published), (2) trained and tested our TeX-Net to demonstrate the efficacy of TeX vision on real-world level scenes, and (3) compared the machine-learning performances, especially detection and ranging, based on our TeX vision and the traditional thermal vision. **Our new results (see Extended Data Figs. 1, 2, 6, 8 and 11 in the new version) clearly show the efficacy of HADAR in real-world level complicated scenes.** In the following, we will further explain our results to demonstrate HADAR efficacy.

1. To generate the synthesized database, we used Monte Carlo path tracing (Planck’s law + Kirchoff’s law + Blender Cycles) and designed a city block scene to mimic a real-world self-driving task. This city block scene is rendered with multiple scattering cutoff of $l = 4$ (i.e., ray depth = 4), which is commonly adopted for real-world level image quality especially for low-reflection materials. In this city block scene, there are 21 different material categories ($M=20$ for the material library, robots share the same material with car logo). For comparison, we note that state-of-the-art semantic segmentation of optical imaging in the literature has similar number of categories. For example, the pre-trained DANET [1] was trained to segment 19 categories, while the CityScapes dataset (<https://www.cityscapes-dataset.com/>) has 30 classes for segmentation. One key difference of our approach is **physics-driven semantic segmentation**. State-of-the-art semantic segmentation focuses on object level distinctions within the scene (e.g., car, road, pedestrian, etc.). However, our approach for TeX decomposition focuses on materials and hence exploits the unique thermal signatures at the physical-component level (e.g., car paint, window, headlights, tire, etc.), which is more advanced, see the following figure.

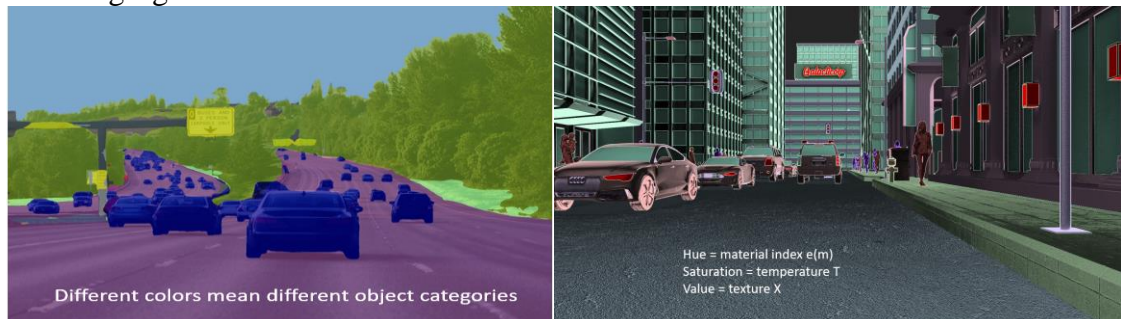


Figure 8. Left: a sample of the state-of-the-art semantic segmentation in the object level (ADE20K), with 6 categories. Right: a sample TeX vision in our HADAR database, **with 20 categories**. This comparison is to show the real-world complexity of our dataset.

In our city block dataset, there are multiple pedestrians, including men, women, kids, the elders, and robots, to mimic a future scene. We believe our database presents a convincing platform to test HADAR efficacy (TeX decomposition and TeX vision).

2. Based on our synthesized database, the machine learning performances of our TeX-Net (see Extended Data Fig.1) shows the applicability of TeX decomposition and TeX vision on complicated scenes. This demonstrates HADAR efficacy on synthesized real-world scenes. The comparison of TeX-Net output with the ground truth TeX vision is given in Fig.S12 in the Supple. Info., as cited below.

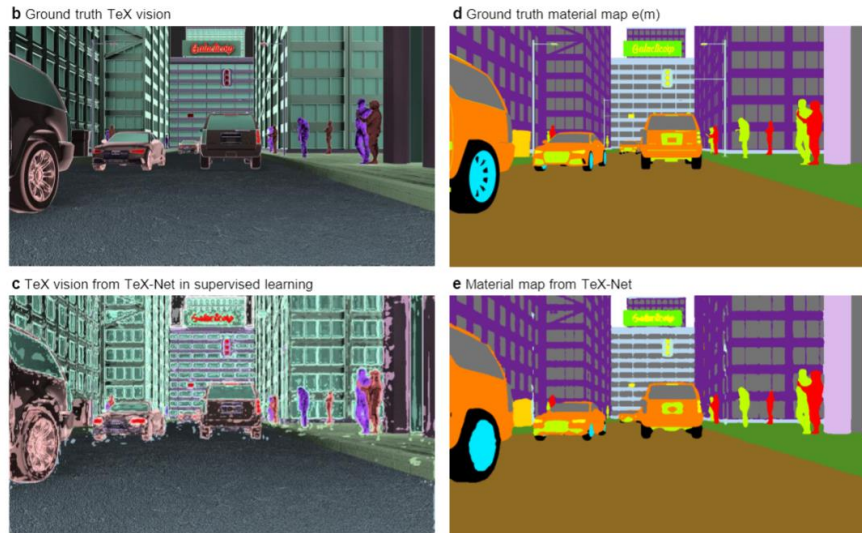


Fig.9 Comparisons of TeX-Net output with the ground truth show that TeX-Net is indeed able to perform TeX decomposition. Small prediction errors in temperature lead to texture error in brightness, and hence there are some noisy spots observed in c. This can be improved by imposing sophisticated smooth constraint on temperature and harder training.

3. In this paper, we demonstrate a few typical examples, i.e., people detection with HOG+SVM, semantic segmentation with DANet, and ranging with DeepPruner. Performance comparisons between TeX vision and traditional thermal vision clearly indicates HADAR efficacy, as shown in Extended Data Figs. 5, 6, 8, and Fig. S17 of the Supple. Info. We briefly cite the results as below for the reviewer's convenience. We note that HADAR requires multi-spectral information to output TeX vision. **In the proposed concept of TeX vision, the scene is captured with physical attributes being represented by hue (material index), saturation (temperature) and value (texture).** This novel representation has physical context which is not present in the output of optical cameras (RGB vision), conventional IR thermal cameras (panchromatic thermal vision), or LiDAR (point cloud). Subsequent machine learning algorithms in computer vision regarding stereo matching, optical flow, scene flow, semantic segmentation, etc. that are previously based on RGB vision, thermal vision or point cloud can be adapted to TeX vision. Developing new algorithms exploiting TeX vision presents a new research frontier and we plan to pursue multiple avenues in future studies.

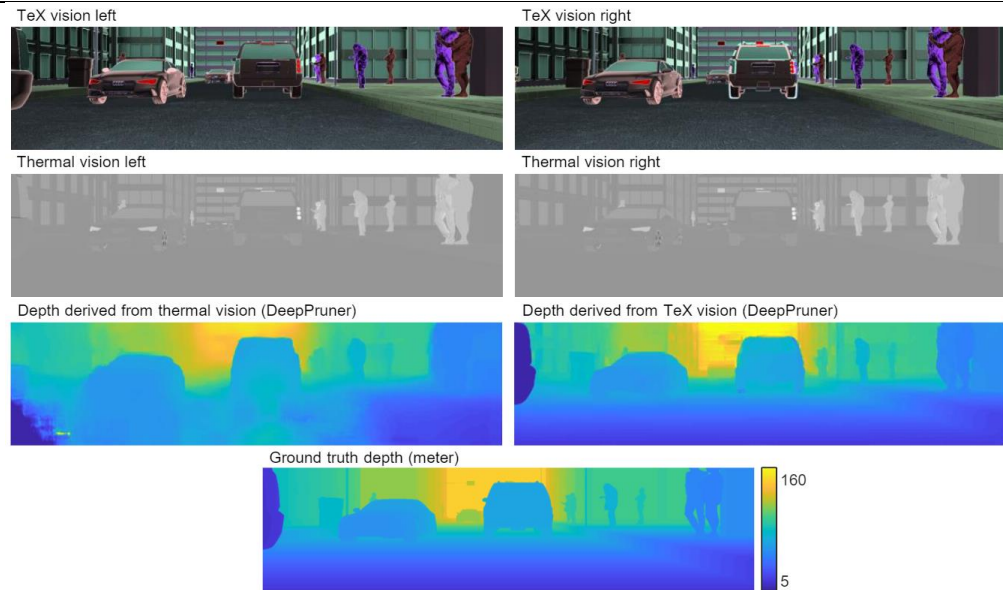


Fig.10 ‘TeX vision + AI’ beats the state-of-the-art ‘thermal vision + AI’ in ranging, showing HADAR efficacy in complicated scenes.

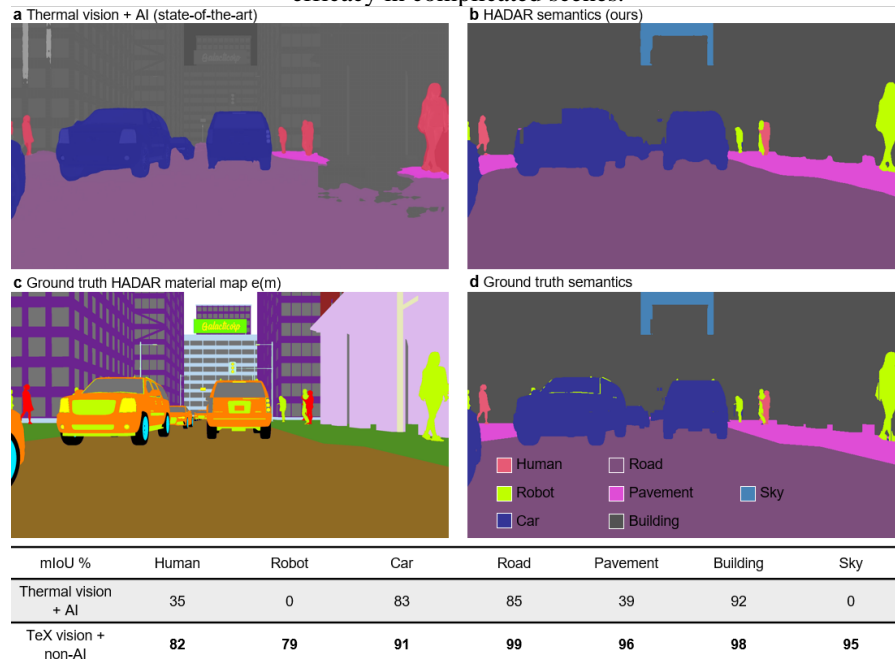


Fig.11 HADAR semantic segmentation based on TeX vision beats the state-of-the-art ‘thermal vision + AI’ segmentation, showing HADAR efficacy in complicated scenes.

- Finally, we want to comment on the development of a real-world HADAR dataset. We note that a real-world experimental hyperspectral LWIR database with ground truth TeX vision and depth is not available. The procedures to build an experimental database are basically the same as what we have done in the outdoor experiments, and they can be briefly summarized as the following: (1) collecting left and right hyperspectral data

cubes; (2) collecting the corresponding material library; (3) solving TeX as the ground truth by least-squares estimators; and (4) collecting ground truth depth with LiDAR. In this first paper, we have provided a synthetic database and shown two proof-of-concept experiments in summer and winter. This forms a foundation to develop a real-world HADAR dataset. The two key steps are outlined below.

Collecting material library: In our current experiments, we used a subset of the NASA JPL ECOSTRESS spectral library as our material library. This library is for Spaceborne applications, not self-driving cars. Consequently, there are many materials (e.g., human skin, hair, clothes, and tires of cars) common in daily life but missing in the library, since they are rare to be seen from the space. Instead, we have to use other similar materials to approximate the spectral emissivity. Note that TeX vision requires spectrally resolved emissivity different from existing panchromatic thermal vision where emissivity is approximated as a single number (i.e. $e(\lambda)$ vs $e=\text{constant}$). We did observe residual errors in our results due to the mismatch of emissivities used in the algorithm with respect to the actual emissivities. The error manifests in the texture map especially around boundaries in Fig.5 and Extended Data Fig.10; if the material library is perfectly known, one can recover texture as accurate as Extended Data Fig.2c. We intend to follow the same procedures as the JPL database [2] to generate a standard material database for self-driving applications. Handheld spectrometers can be used to collect the material library instead of bench-top spectrometers used in [2]. Building a material library includes spectrometer calibration, sample preparation, measurement, error analysis, and especially cross analysis with the JPL library for shared materials.

Spectral resolution of thermal image: Secondly, to distinguish larger number of materials in the HADAR material library requires better spectral resolution, in our case, more spectral filters. However, research and commercialization of LWIR spectral filters currently lag behind visible-light filters. Especially in the COVID period, the 10 filters we used are the only significantly independent filters available in stock from Spectrogon, an industry-leading company providing LWIR filters. We are fabricating custom spectral metamaterial filters to enhance the resolution. Alternatively, using grating-/interferometer-based hyperspectral imagers could be another solution.

Our group has extensive preliminary work on those various aspects, and we are confident that the above two factors can be overcome in the near future. We do provide the research community with this first set of HADAR data collected in Indiana. However, building a standard material library or building a hyperspectral imager are independent projects beyond the scope of this first paper on HADAR.

Synthesized database is commonly used in the literature, for example, the Scene Flow dataset (<https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html>) for depth estimation and segmentation, and the MPI Sintel benchmark for optical flow (https://ps.is.tuebingen.mpg.de/research_projects/mpi-sintel-flow). Synthesized data, especially by Monte Carlo path tracing, has the real-world image quality and complexity, and has perfect ground truth and calibrations. We believe this dataset will motivate the research community to pursue multiple avenues related to TeX vision and HADAR.

	<p>Reference(s):</p> <p>[1] Fu, Jun, et al. "Dual attention network for scene segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.</p> <p>[2] A. Baldrige, et al, "The aster spectral library version 2.0", Remote Sens. Environ. 113, 711 (2009).</p>
C4.2	<p>How do the authors do HADAR ranging and stereo vision? A more in-depth description of this process would aid clarity of the paper and understanding of the reader.</p>
R4.2	<p>We note that we use all three coupled physical attributes of TeX vision: temperature, emissivity, and texture for computing the range. TeX vision exploits spectral information in infrared heat radiation along with the HADAR constitutive equation (Eq 1) to separate intrinsic and extrinsic thermal photons. This physics-driven approach gives rise to three information channels which we represent as hue (emissivity), saturation (temperature) and value (texture). This is fundamentally different from optical cameras which output RGB vision. HADAR ranging is based on stereo matching of left and right TeX vision images, like traditional stereo matching on RGB vision images. Thus, the large number of neural network architectures used in optical vision tasks can be adopted to TeX vision in the near future.</p> <p>(1). We apologize that in the previous version of Fig.1b, it was confusing to put depth before TeX vision. In this new version, we revised it to the following.</p> <div data-bbox="300 940 1412 1207" data-label="Diagram"> <p>The diagram, labeled 'b', illustrates the HADAR process. It starts with a 'Ghosting effect (TeX degeneracy)' showing two overlapping human figures. This leads to a 'HADAR' sensor that captures a 'Hyperspectral heat cube'. This cube is processed by a 'TeX-Net' neural network. The output is decomposed into three channels: 'Temperature' (T), 'Emissivity' (e), and 'Texture' (X). These channels are combined to produce 'TeX vision', which shows the two figures with distinct color coding (one red, one green). This TeX vision is then used for 'HADAR perception', resulting in 'Enhanced depth' and semantic segmentation, where the 'Metallic robot' is highlighted in red and the 'Human body' in blue.</p> </div> <p>Fig.12 HADAR outputs TeX vision. Detection, semantic segmentation, ranging, etc are based on TeX vision. Depth is computed after TeX vision</p> <p>(2). To make our approach clearer, we have directly stated in the introduction (last sentence, 1st page) that:</p> <p><i>“Our demonstrations of HADAR include detection and ranging based on TeX vision, for both real-world level HADAR database and outdoor experiments.”</i></p> <p>(3). In HADAR ranging section of the main text and Fig.4, stereo matching is used only for the scattering signal, $(1-e)X$, to show the importance of texture in ranging and compare with optical imaging. This is because the main text focuses on describing the fundamental limits of HADAR, and it is more intuitive to illustrate the ranging error bound with only the scattering signal. However, we emphasize that our fundamental bound on ranging error is universal and applies to all kinds of images, including the TeX vision, X-type texture, the scattering signal, or even optical images. We have added these explanations in the HADAR ranging section to make the logic flow more fluent:</p> <p><i>“...To show the importance of texture in ranging and compare with optical imaging, here we focus on the scattering signal that can be reconstructed through TeX decomposition...”</i></p>

(4). Moreover, HADAR ranging based on TeX vision is also explicitly demonstrated in Extended Data Fig.6. Stereo matching is performed with DeepPruner pre-trained on the KITTI dataset.

C4.3 Also, what is meant by ~100 x accuracy in ranging? Does this apply to all scenarios rather than specific cases?

R4.3 The previously mentioned 100-fold improvement in ranging accuracy is for one single line in the specially designed scene shown in Fig.4 of the main text. It does not apply to all scenarios. In that special scene in Fig.4, ranging accuracy enhancement is between 1~100 at different image parts. We focused on one single line in Fig.4 in order to **verify our fundamental bound** of HADAR ranging.

Ranging results on general scenes like the city block dataset are given in the Extended Data Fig.6, as cited below. We have observed about **67 folds** mean disparity error suppression of HADAR ranging with respect to thermal ranging for the street region and **26 folds** mean disparity error suppression for the entire image. Note that disparity error is proportional to ranging error. In the new version, statistics have been analyzed across multiple frames.

		Density	Mean error (px)	Accuracy (%)			
				$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 10$
Thermal	Street	0.5	99.74	4.45	12.65	17.01	22.29
vision + AI	Entire image	1	55.32	31.54	45.99	50.36	54.98
TeX vision	Street	0.5	1.49	24.74	96.94	98.77	99.66
+ AI	Entire image	1	2.09	42.20	93.59	96.64	98.45

For the fundamental bound of ranging accuracy, the improvement factor is related to $\eta = \sigma_{c,t}^2/\sigma_{c,h}^2 = J_{x,h}^0/J_{x,t}^0$, where $\sigma_{c,h}^2$ is the photonic correspondence uncertainty with spectral resolution (HADAR), $\sigma_{c,t}^2$ is the photonic correspondence uncertainty without spectral resolution (thermal vision), $J_{x,h}^0$ is the Fisher information with spectral resolution, and $J_{x,t}^0$ is the Fisher information without spectral resolution, as given in Tab. S4 of the Supple. Info. **This improvement factor η is scene dependent but is always greater than or equal to 1** (between 1 to ∞). Relevant discussions are given in page 33 of the Supple. Info. as cited below.

Eq. (S43) recovers Rayleigh's limit. We now briefly prove that the Fisher information for HADAR ranging with spectral resolution is more than Fisher information for panchromatic thermal imaging. In the mathematical expression of the Fisher information for HADAR in Tab. S4, the spectral information is squared before integral, which prevents destruction of the spectrally resolved Fisher information from contributions of opposite signs. This leads to a larger Fisher information J_x^0 and a smaller photonic correspondence uncertainty σ_c . Mathematically, $\int \frac{(\partial_x p_{x\nu})^2}{p_{x\nu}} d\nu - \frac{(\int \partial_x p_{x\nu} d\nu)^2}{\int p_{x\nu} d\nu}$ can be manipulated into a square form, $(*)^2 \geq 0$, $*$ being a certain expression, and hence it proves that the Fisher information is larger with spectral resolution. More importantly, by breaking the TeX degeneracy, HADAR can support sophisticated priors like sparsity or smoothness to further remove unknowns in the parameter set $\{m_\alpha, T_\alpha, V_\alpha\}$, suppressing ranging error toward a lower bound, $J_x^0 \leq \iint_\Omega \frac{(\partial_x b_{x\nu})^2}{b_{x\nu}} + \frac{(\partial_x k_{x\nu})^2}{k_{x\nu}} dsd\nu$, with $b_{x\nu} \equiv \tilde{S}_{x\nu}^0 / \iint_\Omega S_{x\nu} dsd\nu$ and $k_{x\nu} = p_{x\nu} - b_{x\nu}$. Here, $\tilde{S}_{x\nu}^0$ is the direct emission.

	<p>In order not to mislead readers, we have revised relevant statements: <the second sentence in HADAR ranging section, page 3> “...We prove a transformative 2-orders-of-magnitude accuracy improvement in depth estimation with HADAR ranging, as compared with existing thermal ranging. We demonstrate depth accuracy improvement of HADAR ranging up to two orders of magnitude compared with existing thermal ranging...” <last sentence of the first paragraph in HADAR ranging section> “...the absolute ranging error (cyan data points in insets) with respect to the ground truth along white dashed lines shows $\sim 100\times$ accuracy improvement in HADAR (<i>The improvement is scene dependent. See Extended Data Fig. 6 for general scenes and Fig. 14 for experimental evidence of the advantage of HADAR ranging</i>)”</p>
C5	<p>As per the authors, the HADAR framework requires a hyperspectral cube to input to the NN. Despite such promising performance of HADAR, its applicability on real-world datasets might be limited as most of the real-world thermal datasets are only available based on temperatures. Do authors intend to extend the HADAR estimation to extrapolate the rest of the information based on this limitation in information availability for more applicability?</p>
R5	<p>We agree with the reviewer that existing real-world thermal datasets are only based on panchromatic radiance (commonly treated as temperature) without spectral resolution, while spectral resolution in hyperspectral data cubes is very essential to HADAR to obtain the TeX vision. We have created the first stereo-hyperspectral Cityblock dataset to aid the community in exploring fundamental bounds of TeX vision and HADAR. Specific aspects of HADAR can be extended to existing thermal datasets without spectral resolution in the following way</p> <ol style="list-style-type: none"> 1. The fundamental bound on ranging accuracy in our theory is universal, applicable to various kinds of data including hyperspectral data cubes, RGB images, and even panchromatic grayscale images. The bound is fully given in Sec.SIIC of the Supple. Info. as cited below, $\sqrt{N}\delta z \geq \frac{z^2}{bf} \sqrt{2(1+\gamma)(\sigma_c^2 + \sigma_d^2)}, \quad (\text{S43})$ <p>where $\sigma_c = 1/J_x^0$ is the photonic correspondence uncertainty given by single-photon Fisher information J_x^0. In this new version of manuscript, we have provided in Tab. S4 of the Supple. Info. the Fisher information J_x^0 about point-source location for both thermal data without spectral resolution and hyperspectral data cubes with spectral resolution. Furthermore, as can be seen in insets of Fig.4 in the main text, the theoretical predictions (red curves) of our HADAR ranging bounds are consistent with numeric experiments (blue dots), both with and without spectral resolution. Even for practical imaging without spectral resolution where the exact photon number is unknown, we also provided the corresponding Fisher information about window position as cited below. The associated Cramer-Rao bound (inverse of Fisher information) bounds the ranging accuracy.</p> $J_x = \langle \partial_x \log \bar{\mathcal{P}}(n) \cdot \partial_x \log \bar{\mathcal{P}}(n) \rangle = \frac{(\partial_x N_{iq})^2}{N_{iq} + \sigma^2}. \quad (\text{S37})$

Our fundamental bound on ranging accuracy is applicable to existing thermal datasets. In our current paper, we have only compared with our own dataset but will extend to other thermal datasets in the future.

2. The fundamental bound of HADAR identifiability (material estimation) cannot be extended to thermal datasets without spectral resolution. Spectral thermal signatures are key to material estimation, while panchromatic thermal imaging loses spectral information.
3. As an extension of our theory, we propose pseudo-TeX vision for existing panchromatic thermal imaging datasets, to extend the practical application regime of TeX vision. For high emissivity objects, thermal image is widely approximated as the temperature contrast. Standard thermal cameras can do the inverse transform and provide a rough estimate of the temperature. Therefore, we use the thermal image itself to approximate temperature T . Secondly, existing semantic segmentation based on thermal vision can extract spatial patterns (geometry) from thermal images and estimate semantic categories. We use it to approximate material category $e(m)$. Thirdly, AGC (automatic gain control) can improve visual contrast, maximizing the usage of residual texture in sensor data. We use it to approximate texture X . In doing so, we get an approximation to the three attributes of T , e and X (pseudo-TeX vision). Please see the following Fig.14 for a sample thermal image in the FLIR thermal dataset without spectral resolution (<https://www.flir.com/oem/adas/adas-dataset-form/>). We emphasize that spectral resolution is crucial to accurate temperature estimation, material classification as well as texture recovery.

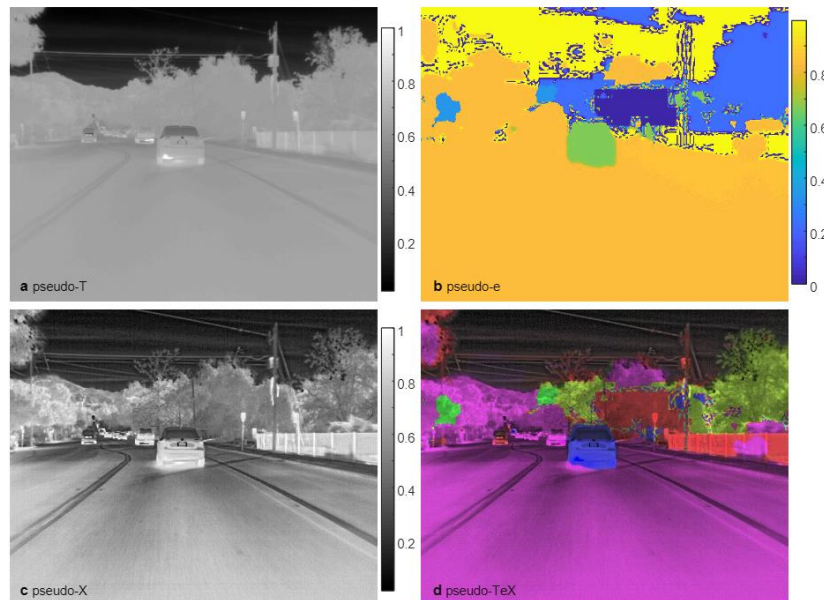


Fig.13 Pseudo-TeX vision for a sample thermal image in the FLIR thermal dataset without spectral resolution (<https://www.flir.com/oem/adas/adas-dataset-form/>).

We emphasize that the fundamental bounds are not improved in pseudo-TeX vision. Pseudo-TeX vision uses information of different levels (spatial pattern, rough

temperature, and weak variation) to extrapolate the material and geometry information and might find applications, for example, in practical ranging, as shown below. Pseudo-TeX vision has been added in Sec.SIID of the Supple. Info.

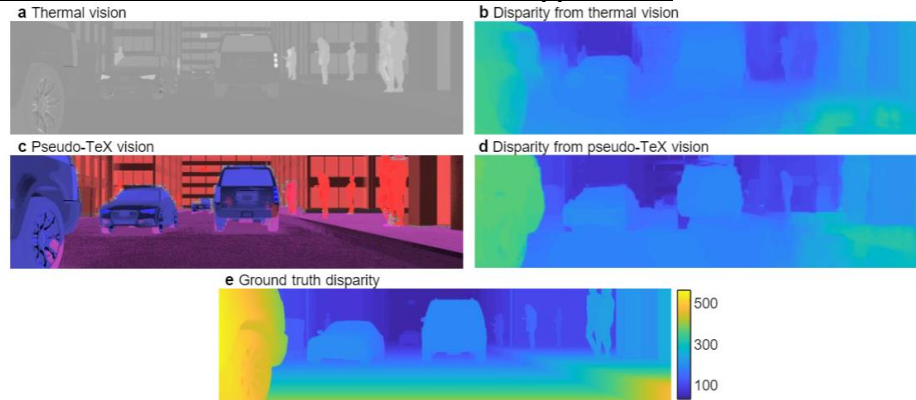


Fig.14 Disparity in stereo matching based on raw thermal vision and pseudo-TeX vision. This scene is one frame of the city block dataset.

At last, we want to add that (1) hyperspectral imaging technology is developing fast in recent years. Our group has worked on developing hyperspectral imagers for years and we know that other than the filter wheel approach, there are color mosaic sensors [1], spatial separation approaches (prism-based push broom [2], grating/AOTF [3]), and interference approaches (Michelson interferometer [4], Fabry-Perot cavity [5]) to obtain hyperspectral data cubes. (2) We have built and released the 1st LWIR stereo-hyperspectral HADAR dataset by exploiting Planck's law and Kirchhoff's law in Blender Cycles. With continuous efforts in the near future, we are generating larger and larger LWIR hyperspectral datasets, both synthetic and experimental, for the scientific community to develop HADAR algorithms.

Reference(s):

[1] Bao, Jie, and Mounji G. Bawendi. "A colloidal quantum dot spectrometer." *Nature* 523.7558 (2015): 67-70.
 [2] Mouroulis, Pantazis, Robert O. Green, and Thomas G. Chrien. "Design of pushbroom imaging spectrometers for optimum recovery of spectroscopic and spatial information." *Applied Optics* 39.13 (2000): 2210-2220.
 [3] Gupta, Neelam, Rachid Dahmani, and Steven J. Choy. "Acousto-optic tunable filter-based visible-to-near-infrared spectropolarimetric imager." *Optical Engineering* 41.5 (2002): 1033-1038.
 [4] Potter, Kimberlee, et al. "Imaging of collagen and proteoglycan in cartilage sections using Fourier transform infrared spectral imaging." *Arthritis & Rheumatism* 44.4 (2001): 846-855.
 [5] Lucey, Paul G., et al. "A compact Fourier transform imaging spectrometer employing a variable gap Fabry-Perot interferometer." *Next-Generation Spectroscopic Technologies VII*. Vol. 9101. International Society for Optics and Photonics, 2014.

C6	The NN devised for TeX decomposition is a flattened fully connected layer/1D CNN, which results in the loss of spatial-spectral information during the inferences. How much of that is evident in current results? Quantification of results is missing and needs to be added to make the argument for HADAR application more compelling. Will a 3D CNN based on spatial-spectral info be more useful for better performances?
R6	We thank the reviewer for the suggestion of using spatial-spectral information. In the previous version, we were considering a simplest model that decomposes TeX attributes pixel per pixel, using spectral information but ignoring spatial information. That previous simplest model (3-

layer 1D CNN) is *unable* to process complicated scenes. Instead, we have now used 3D CNN based on spatial-spectral information in this new version, and we have observed better performances of TeX decomposition utilizing both spatial patterns and spectral signatures. Explicitly, to improve the quality of this manuscript, (1) we have built and released the 1st HADAR Cityblock stereo-hyperspectral dataset with real-world level complicated scenes (https://drive.google.com/drive/folders/1da2Uh5t_QOy-MrWxhkJJw3MueNxsuVtn?usp=sharing), and (2) we have adopted spectral and pyramid attention layers in the UNet model [1] and proposed the TeX-Net for TeX decomposition, as cited below.

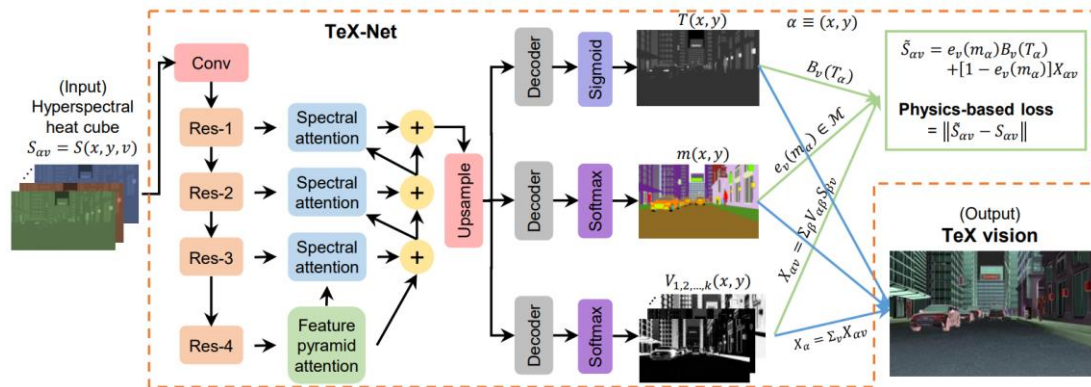


Fig.15 Architecture of TeX-Net for TeX decomposition. The input to TeX-Net is the hyperspectral heat cube. The output is the TeX vision. Loss function to train TeX-Net is defined on the reconstructed heat cube, which is based on physics models of blackbody radiation, the material library, and the mathematical structure of texture X. This figure is given in Extended Data Fig.1.

TeX-Net has built-in 3D CNN's and is using spatial-spectral information. The training and performance of TeX-Net is given in Sec.SIIIA of the Supple. Info., as cited below.

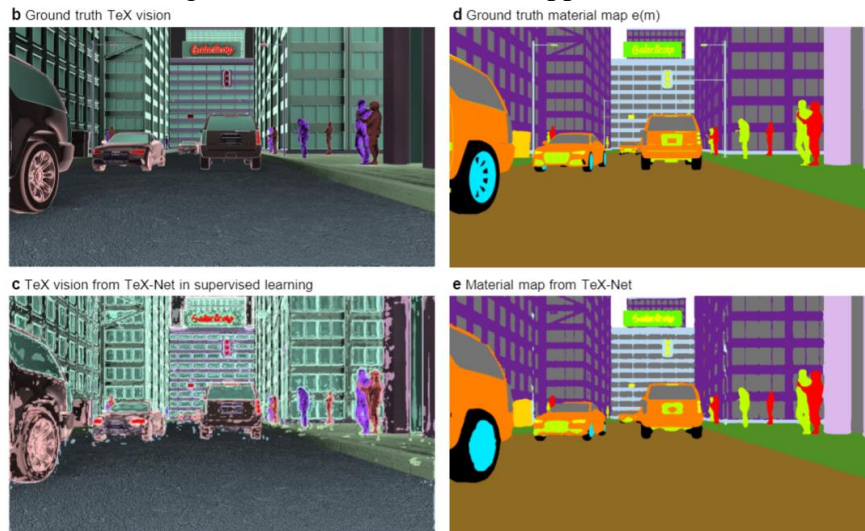


Fig.16 Comparisons of TeX-Net output with the ground truth show that TeX-Net is indeed able to do TeX decomposition. TeX-Net is trained with supervised learning. Small prediction errors in temperature lead to texture error in brightness, and hence there are some noisy spots observed in c. This can be improved by imposing sophisticated smooth constraint on temperature and harder training.

	<p>Reference(s): [1] Li, Hanchao, et al. "Pyramid attention network for semantic segmentation." arXiv preprint arXiv:1805.10180 (2018).</p>
C7	<p>It is not clear if the authors can identify pixels corresponding to different material from 1D CNN. Is the material identification done at a global image-level or pixel-by-pixel basis? Would 1D CNN suffice for the task if it is done pixel by a pixel basis? Something like UNET based models would be apt for such tasks.</p>
R7	<p>As explained above in Comment & Reply 6 (R6), the previous simplest model of 1D CNN is working on a pixel-by-pixel basis. That simplest model was a 3-layer CNN for material classification followed by analytical expressions to solve temperature T and texture X. The previous model can identify materials per pixel for our simple synthesized scenes shown in the paper but cannot work on real-world level complicated scenes. We thank the reviewer again for the suggestion of using 3D CNN, spatial-spectral information, and UNET-based models. <u>In this new version, we have developed TeX-Net for TeX decomposition.</u> TeX-Net has built-in 3D CNN's and has both spatial and spectral attention layers. TeX-Net is inspired by the Pyramid Attention Network [1] which uses UNet. TeX-Net identifies material for each pixel at a global-image level, using both spectral and spatial information.</p> <p>Reference(s): [1] Li, Hanchao, et al. "Pyramid attention network for semantic segmentation." arXiv preprint arXiv:1805.10180 (2018).</p>
C8	<p>In the supplementary material for HADAR estimation theory, in the last sentence of the section, Theory of Texture, the authors mention that HADAR texture is equivalent to grayscale imaging in daylight, which the reviewer thinks is False as grayscale imaging posses greater qualitative and quantitative texture information than that of HADAR as inferred by the results currently presented in the paper. Improving results to substantiate this claim or else removing it is suggested.</p>
R8	<p>We regret that we made an unclear statement in the previous version saying ‘HADAR texture is equivalent to grayscale imaging in daylight’. <u>We have removed the above unclear sentence in the new version.</u></p> <p>The message we tried to convey is the following. Grayscale optical imaging in daylight which possesses textures uses the scattered light signal from objects. Thermal imaging suffers from the ghosting effect because the scattered signal is immersed in strong direct heat emission. HADAR reconstructs the scattered signal and hence recovers the thermal texture. We agree that materials’ response to light (spectral features of emissivity) in the visible-light spectrum is different from that in the thermal infrared spectrum. Therefore, HADAR has different textures than grayscale imaging in daylight even though HADAR reconstructs the scattered signal.</p> <p>To make our argument clearer and to improve the quality of our results in the new version,</p> <ol style="list-style-type: none"> 1. We have expanded the Supple. Info. to explain in detail the theory of thermal textures, see Sec.SID for more details. Explicitly, we have explained the above arguments in one of the paragraphs as cited below.

Generally, what is captured by detectors is a mixture of all three types of thermal textures. In traditional thermal imaging, T-type thermal texture dominates, and e-type and X-type textures are weak since $e_{av} \approx 1$. This is why thermal imaging is widely taken as the temperature contrast. T-type thermal texture could give the contour of objects that have different temperature with the background but is poor in details, exhibiting the ‘ghosting effect’. In optical imaging under solar illumination, only the scattering term $(1 - e_{av})X_{av}$ in Eq. (S2) is recorded. Existing object detection, semantic segmentation and stereo depth estimation based on optical images, make use of the e-type and X-type textures. Therefore, to overcome the ‘ghosting effect’ in thermal imaging and implement HADAR with comparable performance to optical detection and ranging, the key is recovering and enhancing e-type and X-type textures in the scattering term $(1 - e_{av})X_{av}$, with the help of spectral resolution. However, we remind that materials’ response to light (spectral features of emissivity) in the visible-light spectrum is different with that in the thermal infrared spectrum. This is another difference between thermal imaging and optical imaging.

2. We have quantified textures at both the fundamental level (Fisher information metric) and the visual level (standard deviation metric), see Sec.SIID for more details.
3. We have improved our results about texture recovery, in both outdoor experiments and synthesized dataset, as cited below. The following summer daylight experiment shows HADAR TeX vision in comparison with traditional thermal vision. HADAR texture recovers geometric textures, see the grass. Quantitatively, HADAR TeX vision has a mean texture density of 0.0879 (in standard deviation metric), 4.60 folds more than the state-of-the-art pseudo-color approach (0.0191).

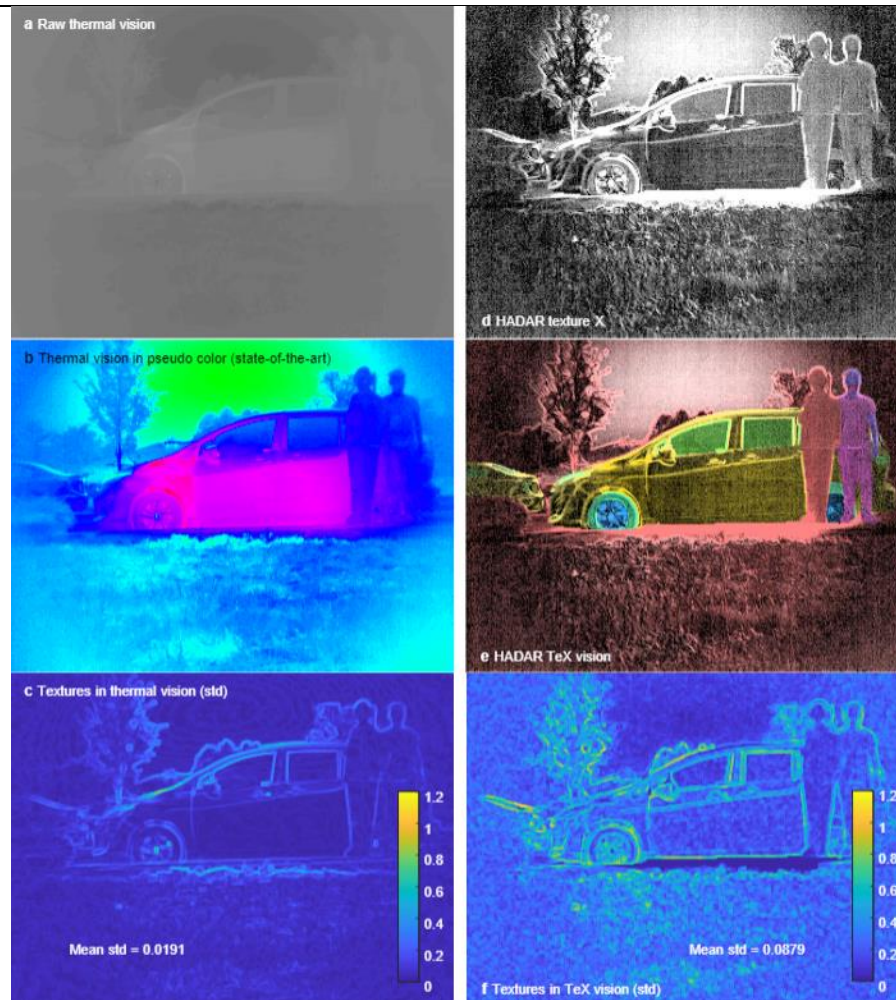


Fig.17 HADAR TeX vision in summer daylight. This figure is a new experiment result given in the Extended Data Fig.10.

We also tested our TeX vision on the city block dataset. The city block dataset clearly shows that TeX vision recovers textures, especially on the road and sidewalks, beating both raw thermal vision and state-of-the-art enhanced thermal vision. Quantitatively, the mean texture density in enhanced thermal vision (standard deviation metric) is 0.0170, while the mean texture density in TeX vision is 0.0788 and is about 4.64 folds larger. This result is given in Extended Data Fig.2. HADAR TeX vision has also been tested on an off-road desert scene in Extended Data Fig.11, where we can see that TeX vision recovers textures. It has the physical appearance comparable to an RGB image as opposed to a conventional thermal image. In the dataset, emissivity in the material library is accurately known. Also, image size of the dataset is 1080*1920, much larger than FLIR A325sc (240*320). These two factors make the TeX vision in dataset much better than proof-of-concept experimental performance of TeX vision.

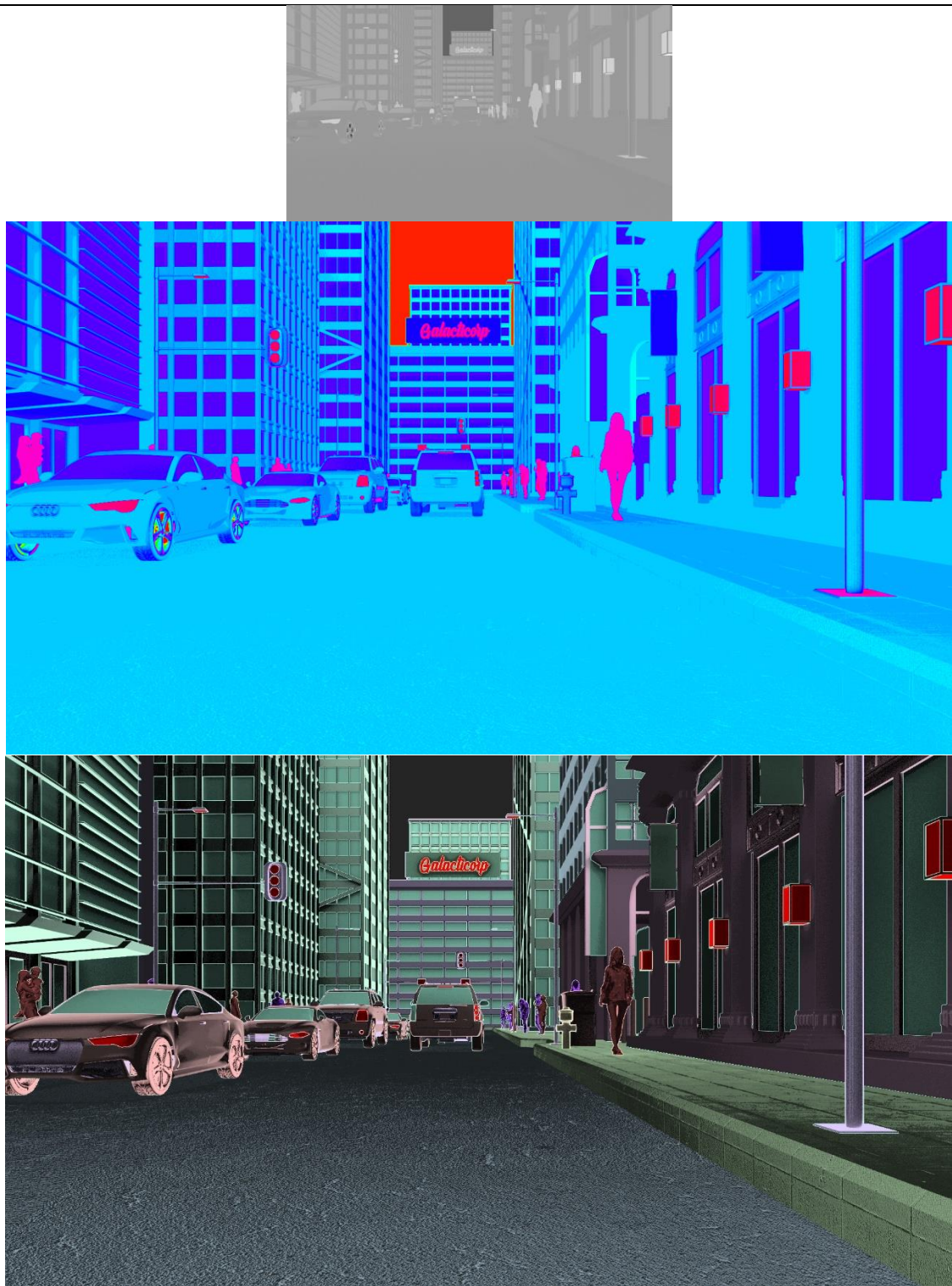
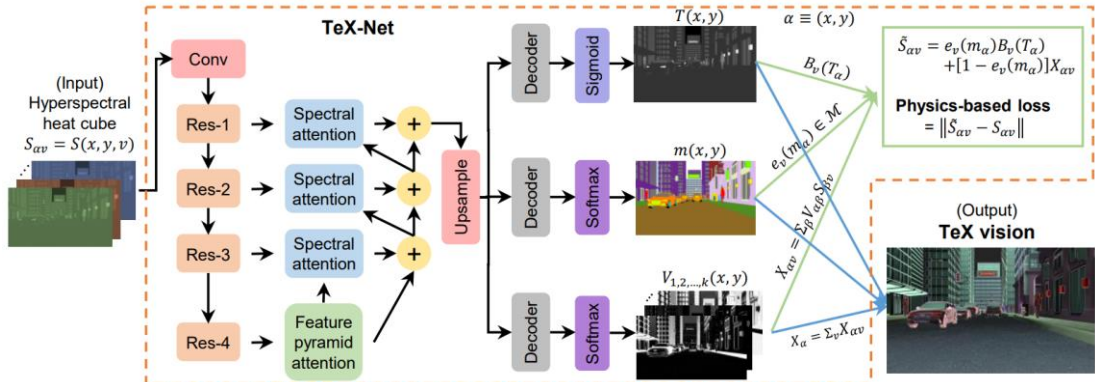


Fig.18 Top: Raw thermal vision with ghosting effect. Middle: State-of-the-art enhanced thermal vision in pseudo color. Bottom: HADAR TeX vision. To better visualize textures, we enlarge TeX vision in comparison with state-of-the-art thermal vision. Texture density figures are not shown here.

C9	<p>To prove the efficacy of the HADAR, the authors estimate material characteristics with NN but use a direct inverse function to estimate T and X. How effective would the application of such inverse function be for real-life applications in comparison to the simulations? Such functions sound promising for theoretical grounds and simulation but involve many constraints and noise factors for real-world application.</p>
R9	<p>We agree with the reviewer about the analytical inverse function for temperature T and texture X. Our previous simplest model of 1D CNN only works for simple synthesized scenes. The previous analytical inverse function is only valid for single-object-single-scattering heat signal model which is unable to process real-world scenes. We used the analytical inverse function to prove the concept of TeX decomposition, but we totally agree that it is essential to demonstrate HADAR efficacy (TeX decomposition) for real-world level complicated scenes. The reviewer is correct that the analytical inverse function suffers from noise as it involves differentiation operations.</p> <p><u>To demonstrate HADAR efficacy in real-world level complicated scenes, in this new version we proposed the TeX-Net for TeX decomposition, as cited below.</u></p>  <p>Fig.19 Architecture of TeX-Net for TeX decomposition. The input to TeX-Net is the hyperspectral heat cube. The output is the TeX vision. Loss function to train TeX-Net is defined on the reconstructed heat cube, which is based on physics models of blackbody radiation, the material library, and the mathematical structure of texture X. This figure is given in Extended Data Fig.1.</p> <p>In un-supervised learning of TeX-Net, the entire network is a mapping from heat cube S to itself. The second part reconstructing \hat{S} from temperature T, material e(m), and thermal lighting factors V is based on the forward analytical function. The first part in the TeX-Net box is technically approximating the inverse functions of T, m, and V. This is a generalization of our previous analytical inverse function. In this revised version, we have provided extensive analysis of TeX vision performance for the multitude of synthesized scenes and two experimental regimes (hot summer day and cold winter night).</p>
C10	<p>For the thermography and semantics experiments, the authors utilize data duplication instead of augmentation to construct a larger dataset. Such methods are not encouraged for DL applications as they might bias the NN towards a certain class. Rather, the application of augmentation techniques is preferred.</p>

R10	<p>We thank the reviewer for the suggestion. <u>In this new version, we have abandoned data duplication completely.</u> In training the TeX-Net, we have built the Cityblock dataset with sufficient training data. In training material classification for human vs. robot identification in Fig.3, we generated sufficient training data with the forward analytical equation of heat signal. Our current approach of TeX-Net uses the input heat cube and down samples it to get the environmental thermal radiation contribution (S_β). S_β is used in reconstructing the texture X, see more details in Sec.SIII of the Supple. Info. Therefore, we have not used data augmentation either in this paper, as it will influence the reconstruction of texture X.</p>
C11	<p>Regarding the applicability of the proposed HADAR setup for application, for dynamic environments such as self-driving vehicles, it would be expensive to have a multispectral acquisition device that can simultaneously acquire the data in the multi-spectrum. Even for just proof of concept in this HADAR framework, the authors change filters via a wheel to retrieve the spectral resolution which may need switching between filters for data acquisition and camera stabilization with such abruptly varying acquisition windows. Such a technique might be problematic in dynamic environments yielding a lot of background noise in TeX decomposition. Suggest adding a paragraph that addresses such current constraints on the real-world application of the method.</p>
R11	<p>We agree with the reviewer that moving HADAR out of lab to the real world will face many practical challenges.</p> <ol style="list-style-type: none"> As suggested by the reviewer, we have added one paragraph in the Methods section --- ‘Prototype HADAR calibration and data collection’ to address data acquisition and functionality-cost tradeoff of HADAR in real-world applications: <i>“...In our proof-of-concept experiments, we used the filter-wheel approach to demonstrate the prototype HADAR. The filter-wheel approach is time consuming but cost effective. HADAR can also be implemented by other approaches such as mosaic sensors, gratings, prisms, interferometers, or Fabry-Perot cavities, depending on the desired spectral resolution, spatial resolution, data acquisition speed, or functionality-cost balance.”</i> We have <u>added a section in Sec.SIIE of the Supple. Info</u> to address camera vibration and non-stationary objects. Generally, camera vibration and dynamic scenes can be described by motion blur and scene flow. We have discussed the applicability of TeX vision and HADAR bounds in the presence of scene flow and motion blur in Sec.SIIE. <p>In the following, we provide in-depth discussions regarding the above constraints, for the reviewer’s information.</p> <ol style="list-style-type: none"> Real-time data acquisition and processing are important to enable high-speed navigation of self-driving cars. It is true that the filter wheel approach for collecting the hyperspectral data cube is time consuming. Switching 10 filters implies it is 10 times slower than thermal imaging. We chose the filter wheel approach only to demonstrate our proof-of-concept experiments, and the reason to use the filter wheel approach is that it is the most cost effective. In the literature, there are many other approaches to collect hyperspectral data cube at a faster speed, for example, the color mosaic sensors [1], spatial separation

	<p>approaches (prism-based push broom [2], grating/AOTF [3]), and interference approaches (Michelson interferometer [4], Fabry-Perot cavity [5]). Color mosaic sensors is as efficient as thermal imaging but sacrifices spatial resolution. Spatial separation and interference approaches have the best spectral resolution but is time consuming, limited in field of view, cumbersome, and expensive. Compared to grating/prism/interferometer approaches, the filter approach is less vulnerable to mechanical vibrations. Improving hyperspectral data acquisition is a hot topic in the scientific community, and fast progress is being made by the community. In processing data, the adoption of TeX-Net instead of least-squares fitting for TeX decomposition enables real-time processing as well as accuracy on complex real-world scenarios.</p> <p>2. Camera vibration, non-stationary objects, and slow data acquisition all lead to motion blur in images. Although our bounds and TeX vision are derived and demonstrated for stationary objects, they are also applicable for non-stationary objects when the motion blur is negligible, that is, when the apparent motion of a point source is within one pixel on the image plane. The apparent motion is given by $\Delta = vtL/r\theta$, where v is the relative transverse speed, t is the exposure time, L is the number of pixels in the horizontal direction, r is the distance of the target, and θ is the field of view. Motion blur is negligible when either the transverse speed is low or the exposure time is short. For example, a target at 30 m away captured by FLIR A325sc ($t < 12$ ms, $L = 320$, $\theta = 50$ degree) equipped on a car driving at 30 mph [$v \leq 30 \sin \frac{\theta}{2}$ mph] will have $\Delta \lesssim 0.8$ and hence the motion blur is negligible. To allow a higher travelling speed, the frame rate of the used camera must be high so that the exposure time is sufficiently short to avoid motion blur, according to the criterion $\Delta < 1$. This criterion, $\Delta < 1$, constrains the applicability of our bounds. Within the criterion, TeX decomposition can be performed for each individual heat cube to get the TeX vision, and detection and ranging are based on TeX vision. We have added the above details in Sec.SIIE of the Supple. Info, with further discussions about the applicability of TeX vision with motion-blur removal.</p> <p>Reference(s):</p> <p>[1] Bao, Jie, and Mounji G. Bawendi. "A colloidal quantum dot spectrometer." <i>Nature</i> 523.7558 (2015): 67-70.</p> <p>[2] Mouroulis, Pantazis, Robert O. Green, and Thomas G. Chrien. "Design of pushbroom imaging spectrometers for optimum recovery of spectroscopic and spatial information." <i>Applied Optics</i> 39.13 (2000): 2210-2220.</p> <p>[3] Gupta, Neelam, Rachid Dahmani, and Steven J. Choy. "Acousto-optic tunable filter-based visible-to-near-infrared spectropolarimetric imager." <i>Optical Engineering</i> 41.5 (2002): 1033-1038.</p> <p>[4] Potter, Kimberlee, et al. "Imaging of collagen and proteoglycan in cartilage sections using Fourier transform infrared spectral imaging." <i>Arthritis & Rheumatism</i> 44.4 (2001): 846-855.</p> <p>[5] Lucey, Paul G., et al. "A compact Fourier transform imaging spectrometer employing a variable gap Fabry-Perot interferometer." <i>Next-Generation Spectroscopic Technologies VII</i>. Vol. 9101. International Society for Optics and Photonics, 2014.</p> <p>[6] A. Baldrige, et al, "The aster spectral library version 2.0", <i>Remote Sens. Environ.</i> 113, 711 (2009).</p>
C12	<p>The authors present a promising and potentially groundbreaking new methodology in their presentation of HADAR. They present significant improvements in performance over other modalities in low light conditions. However, such a modality is not without significant challenges still to be addressed before it can be recognized as a "next step" in computer vision</p>

	<p>applications. From on-fly calibration to the design of acquisition devices pose hardware-level challenges. Besides, the interface of such modules with edge computing devices for real-world applications would be a challenge where the TeX decomposition framework and Task-based frameworks can be easily be deployed in such devices. Another challenge is coming up with a robust library to train the framework as the material properties also change with the environment leading to change in each TeX parameter. <i>Also, the authors don't present the acquisition between cold and hot conditions. How that changes HADAR performance?</i> (will be addressed as C13) Also, for cost-effectiveness, most of the available thermal cameras used in day-to-day life on consumer products are low priced. The applicability of such a multispectral camera for the HADAR application may not be cost-effective and affordable to low-end consumer products and even academic research. The authors make significant projections about the future of this work and have done excellent work in their theory. However, they have neglected to address any of the real and significant challenges that remain before the implementation of such work can take place in a real-world application. Suggest the addition of a section that highlights remaining constraints on the work before it can be presented as a real-world solution. This could take the form of a paragraph in the Discussion or Outlook. The reviewer believes that HADAR brings a lot of potential to the world but as it currently stands, bears significant challenges that need to be carefully addressed before it can replace or substitute the existing modalities in decision making.</p>
R12	<p>We thank the reviewer for the detailed suggestion. We agree that HADAR bears significant challenges that need to be carefully addressed before it can replace or substitute the existing modalities in decision making.</p> <p>In the Reply to last question (R11), we have added two paragraphs to address data acquisition and motion blur. In addition,</p> <ol style="list-style-type: none"> 1. We have also revised our 'Outlook' paragraph in the main text to discuss existing challenges in real-world HADAR application: "We proposed and demonstrated HADAR for fully-passive and physically-aware machine perception. Our shot-noise limits of detection and ranging set the benchmark and call for heat exploitation in the quantum regime where single photon detectors are being developed beyond visible spectral range into the thermal infrared. <u>Practical challenges exist, such as, library collection, on-fly calibration, real-time data acquisition, and functionality-cost optimization.</u> Nevertheless, we believe HADAR will lead to a new chapter in the Fourth Industrial Revolution with applications in autonomous navigation, healthcare, agriculture, wildlife monitoring, geosciences and defense industry." 2. We have also added a paragraph in Methods --- 'pseudo-TeX vision', and a sub-section in Supple. Info., Sec.SIIID, to discuss the applicability of TeX vision in common thermal datasets. See Comment and Reply 5 (R5) for more details. <p>At last, we would like to add that it takes decades in time and enormous efforts from a whole industry for LiDAR to achieve its current breadth of applicability since it was proposed. We hope our work of HADAR could initiate the efforts in industries to boost HADAR improvement.</p>
C13	<p>Also, the authors don't present the acquisition between cold and hot conditions. How that changes HADAR performance?</p>

R13 In the new version, we have added a further experiment in summer daylight (Sep. 2021), in comparison with the previous experiment at a winter night (Dec. 2020).

TeX vision of the summer daylight experiment is given in Extended Data Fig.10, as cited below. In terms of texture recovery performances, the texture density (standard deviation) of TeX vision in summer daylight is about 4.6 folds more than the texture density in state-of-the-art thermal vision. At winter night, the texture density in TeX vision is about 2.9 folds more than the texture density in state-of-the-art thermal vision. This comparison shows that HADAR efficacy is robust on different temperatures. Furthermore, HADAR performance is better in summer daylight (hot temperature) than winter night (cold condition). This is consistent with common thermal experiments, as in hot environments, thermal radiation is stronger and hence the signal-to-noise ratio is higher. Relevant details have been added in the caption of Extended Data Fig.10.

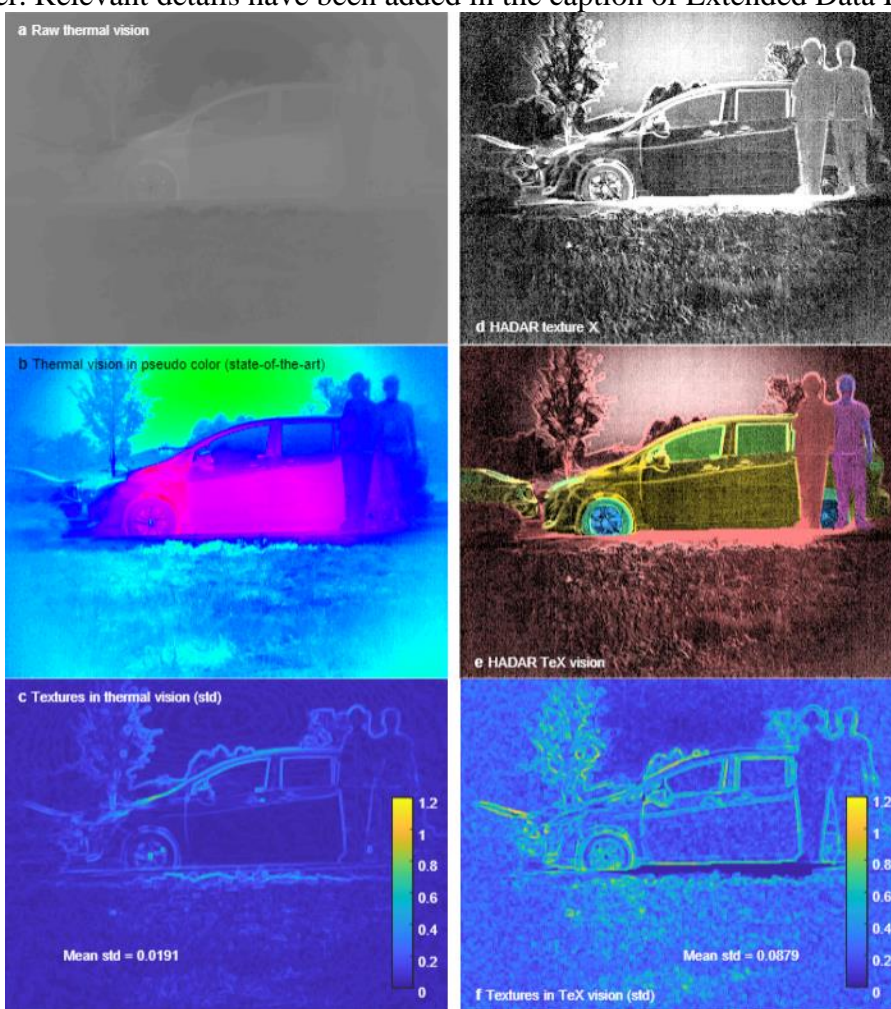
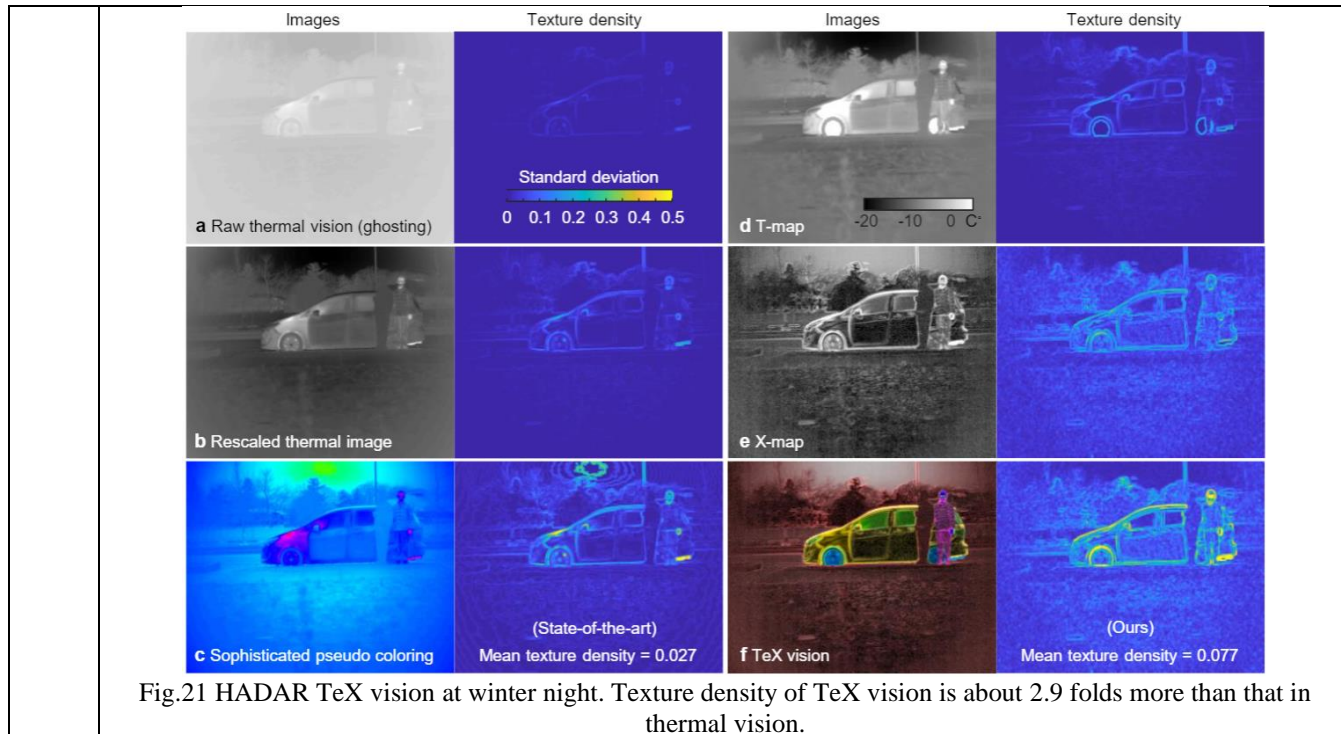


Fig.20 HADAR TeX vision in summer daylight. Texture density of TeX vision is about 4.6 folds more than that in thermal vision.



We thank the reviewer once again for the time and efforts spent to provide us such details comments. With the above changes, we believe the manuscript is now significantly improved and ready for publication.

Cover letter to Reviewer 2

We would like to thank the reviewer for the encouraging response and valuable comments. Here, we list all the major revisions, and we will provide individual replies to each comment from the next page onwards.

Reviewers' main concerns and corresponding major revisions include that

Problem (1):

The previous version of our manuscript needs to be better organized. The derivation of our HADAR theory, details of our machine learning, and details of our HADAR experiments are not provided sufficiently. More examples or explanations are needed.

Revision (1):

We have expanded the Supple. Info., Methods, and Extended Data to provide more details about the theory, machine learning, and experiments, with proper references in the main text.

Problem (2):

The previous version only demonstrated HADAR efficacy for a few simple scenes. HADAR efficacy for complicated scenes is not verified.

Revision (2):

- a) We have built and released the 1st HADAR database with complicated scenes.
- b) We have proposed TeX-Net, based on which, we have clearly shown HADAR efficacy.
- c) We have also added one more experiment in summer daylight to compare HADAR performances on cold and hot conditions, as suggested by other reviewers.

Problem (3):

Quantitative comparisons of HADAR performances (detection and ranging) with the state-of-the-art are incomplete or missing.

Revision (3):

We have made quantitative and qualitative comparisons of machine learning performances based on our TeX vision against the state-of-the-art thermal vision. We show the HADAR advantage for people detection, semantic segmentation, and ranging.

Problem (4):

Textures recovered in HADAR are not quantified and compared to the state-of-the-art approaches.

Revision (4):

We have quantified textures and have made fair comparison with state-of-the-art approaches to show the advantage of HADAR in recovering textures.

Problem (5):

The novelty of TeX decomposition vs. traditional TE separation is not well explained.

Revision (5):

We have explained and emphasized the difference between our TeX decomposition and the traditional TE separation that has been discussed earlier in literature (see R21 for the full reply).

We have also made revisions according to all other comments. Now, we will address each comment sequentially in the following. Notations used in this response include C: Comment, R: Reply, *Italic*: revisions, underline: emphasize.

Reviewer 2	
C0	<p>A. Summary of key results. The state-of-the-art machine perception utilizing active sonar, radar and LiDAR to enhance camera vision is not viable as the number of intelligent agents scales up. Exploiting omnipresent heat signals could be a new frontier for scalable perception. However, objects and their environment constantly emit and scatter thermal radiation leading to textureless images famously known as the ‘ghosting effect’. In this work, the authors proposed a method called HADAR to overcome this ghosting effect by decomposing the heat signal into temperature, emissivity and texture (TeX decomposition). They have developed the HADAR estimation theory and address its shot-noise limits depicting information-theoretical bounds to HADAR-based AI performance. In addition, they have also developed HADAR ranging (depth estimation) that shows an accuracy improvement up to two orders of magnitude compared with existing thermal ranging. They have performed physics-driven semantic segmentation to achieve improved performance against AI-enhanced thermal sensing.</p> <p>B. Originality and Significance This article focuses on the separability of temperature and emissivity from a thermal signal and the use of emissivity profiles for detection and ranging. The separability is discussed earlier in literature and used for various computation assisted tasks. Cramer-Rao bound on the distance for identifiability based on intrinsic properties of a material ascertains quantification of the utility of thermal imaging. Whereas an error bound given on the ranging accuracy with a limitation on photons counts further determines the accuracy of perception. The third component of significance is texture, which is computed from emissivity. Texture in visible light imaging qualifies the identification. TeX decomposition in thermal signal allows fine distinguishable parameters for image processing.</p>
R0	<p>We would like to thank the reviewer for the encouraging response and valuable comments. We have addressed each comment individually below and made major revisions to improve the quality of this manuscript.</p>
C1	<p>C. Data and Methodology There are few comments about Data and Methodology as follows: 1. A schematic diagram of the hardware setup would benefit the readers. Though an image of the hardware setup is provided, it seems insufficient for a scholarly article.</p>
R1	<p>We thank the reviewer for pointing this out. We agree that a better integrated device and a more descriptive picture with 3D schematics are very important for readers to understand HADAR. <u>In this new version, we revised the schematic of hardware setup (previous Extended Data Fig.4, current Extended Data Fig.9), with a real-world picture of our prototype HADAR and corresponding 3D schematics, as cited in the following Fig.1.</u> Furthermore, we also provided the full heat signal models and calibration schematic in the Supple. Info. Schematics of heat signal model are given and explained in Sec.SI. The calibration of our prototype HADAR is fully explained in Sec.SIV.</p>

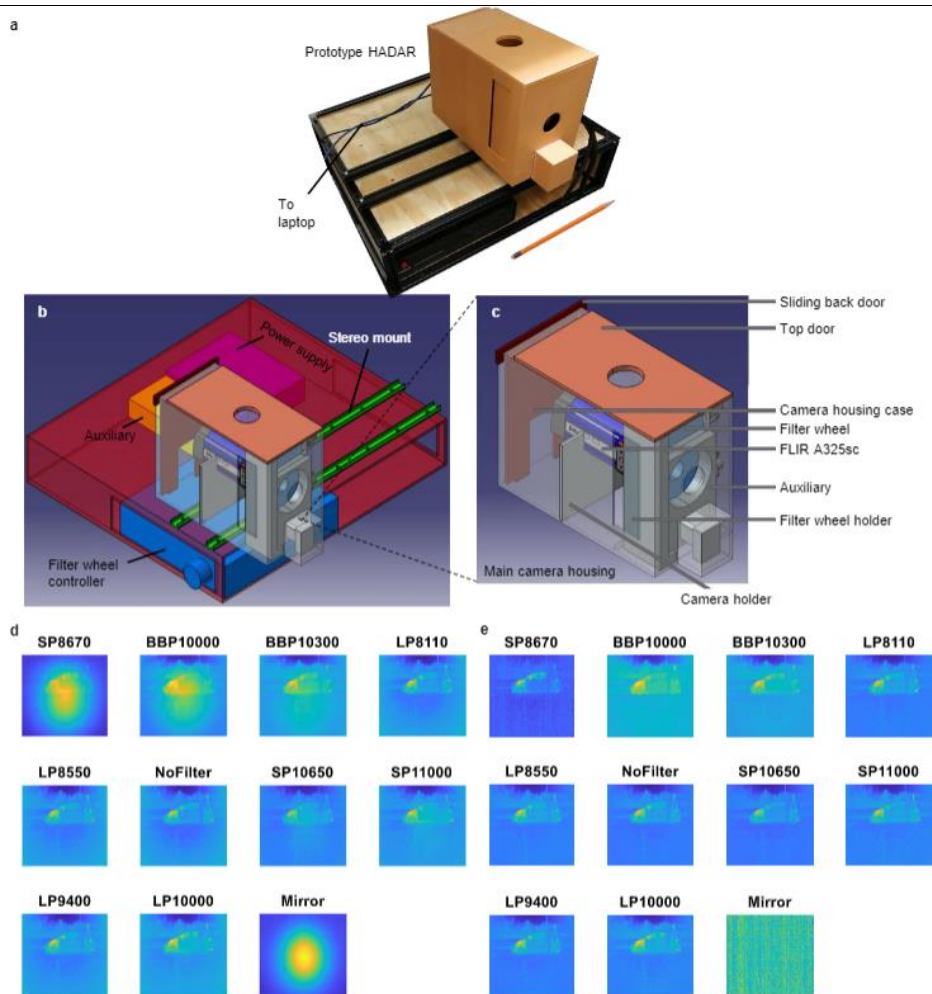


Fig.1 The picture (a) and schematics (b-c) of our prototype HADAR. d-e, HADAR signal before and after calibration of back reflection. This figure is our new Extended Data Fig.9. For more details of calibration, see Sec.SIV of our Supple. Info.

C2	Authors have derived the Cramer-Rao bound for HADAR estimation and the machine learning method but have not provided proof for the Cramer-Rao bound.
R2	<p>We regret that the previous version of the manuscript was over compressed, and the derivation of Cramer-Rao bounds was only briefly explained in the previous Supple. Info. In this new version, <u>we have expanded the Supple. Info to include the full derivations</u> of the fundamental limits for both detection and ranging problems, see Sec.SII of the Supple. Info. The full derivations are based on the unified heat signal model which is explained in Sec.SI of the Supple. Info. (Fig.S2).</p> <ol style="list-style-type: none"> 1. Explicitly, for HADAR identification (material estimation for detection), we used a W-dimensional hyperspace (W being the number of spectral bands) to define the exact detection probability (also known as the recall, or true positive rate; see Fig.S5 and Algorithm 1 in Sec.SIIA). We then developed an analytical theory to derive the detection probability based on Fisher information matrix and the Cramer-Rao bound (see Theorem

	<p>1). The Shannon information or HADAR identifiability of a material within a library was defined by the material’s detection probability given the thermal photons in a scene (see Algorithms 2 and 3). The above full derivation implies that distinguishing two materials is equivalent to estimating the continuous fraction g of a mixture of those two materials. g is continuous and therefore we use a threshold to provide the final discrete estimate of the material index. Based on this equivalency, previously we provided a simpler derivation in previous Supple. Info. We still feel necessary to keep the simpler derivation, which is easier for readers to follow, so we have moved it into the Methods of the paper (Sec. HADAR estimation theory).</p> <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <p>We show that the neural network performance in distinguishing two materials given the thermal signal can achieve the fundamental bound predicted by HADAR estimation theory. Furthermore, this can allow for future comparisons across neural networks to identify the best one which reaches the bound.</p> </div> <p>Numeric Monte Carlo experiments of material classification with machine learning to verify the fundamental limit of HADAR identifiability are given in Fig.3b of the main text and Fig.S6 of the Supple. Info.</p> <p>2. For depth estimation (ranging), we re-interpreted the correspondence problem in computer vision as a window-position estimation problem in estimation theory (see Fig.S8). This allows us to connect the photonic disparity error with the window-position estimation error (see Eqs.S37-41). We then analyzed the physical process of imaging and derived the Cramer-Rao bound of window position uncertainty, which gives the fundamental limit of ranging error (see Eq.S43). Numeric experiments with Monte Carlo path tracing and stereo matching algorithms to verify the fundamental limit of ranging error are given in Fig.4 of the main text.</p>
C3	<p>Authors have selected a single environment for experimentation setup with a few characteristically distinguishable objects (person and cardboard Einstein model). It is evident from the computer vision literature that state-of-the-art artificial intelligence algorithms perform comparatively better in complex scenes. So the efficacy of the proposed method is hard to determine.</p>
R3	<p>We agree with the reviewer that it is necessary to demonstrate HADAR efficacy with complicated scenes and compare with state-of-the-art AI-enhanced thermal vision. To do so, we have (1) built and released the 1st HADAR database with complicated scenes; (2) trained and tested the TeX-Net for TeX decomposition; (3) compared the machine-learning performances based on our TeX vision and the traditional thermal vision; and (4) added one more experiment in a different environment. Our new results (see Extended Data Figs. 1, 2, 6, 8 and 11 in the new version) clearly show the efficacy of HADAR in real-world level complicated scenes. Previously, we used the prototype HADAR in Indiana on a winter night (Dec. 2020) to show 8-material TeX vision (including the background). In the object-level semantic segmentation, previously we only focused on ‘car detection’ and ‘distinguishing the human body and the cardboard Einstein model’, to emphasize the advantage of HADAR over LiDAR and optical cameras.</p>

1. To demonstrate HADAR efficacy in different environments, we have done one more experiment in Indiana, USA, in **summer daylight (Sep. 2021)**, in this new version of the manuscript. This was also suggested by another reviewer to compare HADAR performances on cold and hot conditions. Our results in Extended Data Fig.10 (cited below in Fig.2) of the TeX vision in summer daylight shows consistent TeX capability with the TeX vision on a winter night in Fig. 5. This demonstrates HADAR efficacy of TeX decomposition and TeX vision in different temperature conditions. Furthermore, TeX vision in summer daylight shows better texture recovery. Texture density in TeX vision is about 4.6 folds more than texture density in state-of-the-art enhanced thermal vision. This is generally true that higher temperature would lead to higher signal-to-noise ratio and better performance.

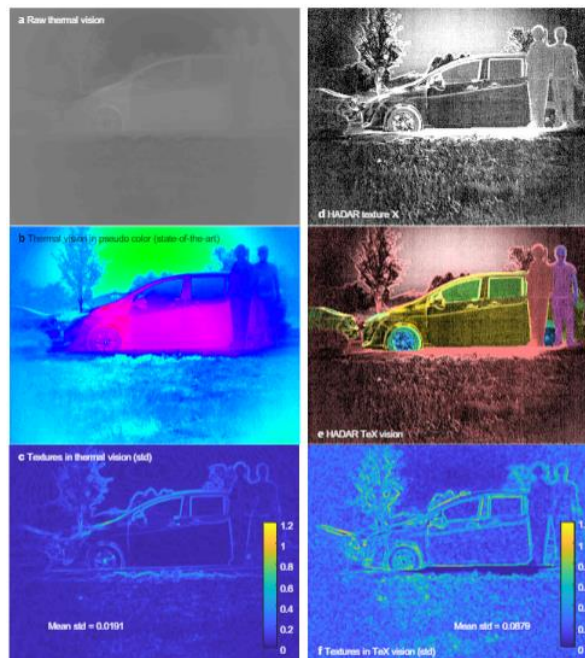


Fig.2 HADAR TeX vision in summer daylight. This figure is a new experiment result given in the Extended Data Fig.10 to show HADAR efficacy on different temperature conditions. For texture quantification, see Sec.SIID of the Supple. Info. for more details.

2. To demonstrate HADAR efficacy in complicated scenes with more objects, we have built a **synthesized LWIR (long-wave infrared) stereo-hyperspectral database** with Monte Carlo path tracing, by exploiting Planck's law and Kirchhoff's law in Blender Cycles renderer. The database has been made available at (https://drive.google.com/drive/folders/1da2Uh5t_QOy-MrWxhkJJw3MueNxsuVtn?usp=sharing), based on which we have tested machine learning performances. We will host it on Github for the research community once the paper is published. To generate the synthesized database, we designed a city block scene to mimic a real-world self-driving task. This city block scene is rendered with multiple scattering cutoff of $l = 4$ (i.e., ray depth = 4), which is commonly adopted for real-world level image quality especially for low-reflection materials. In this city block scene, there are 21 different material categories ($M=20$ for the material library, robots share the same

material with car logo). For comparison, we note that state-of-the-art semantic segmentation of optical imaging in the literature has similar number of categories. For example, the pre-trained DANET [1] was trained to segment 19 categories, while the CityScapes dataset (<https://www.cityscapes-dataset.com/>) has 30 classes for segmentation. One key difference of our approach is **physics-driven semantic segmentation**. State-of-the-art semantic segmentation focuses on object level distinctions within the scene (e.g., car, road, pedestrian, etc.). However, our approach for TeX decomposition focuses on materials and hence exploits the unique thermal signatures at the physical-component level (e.g., car paint, window, headlights, tire, etc.), which is more advanced, see the following figure.

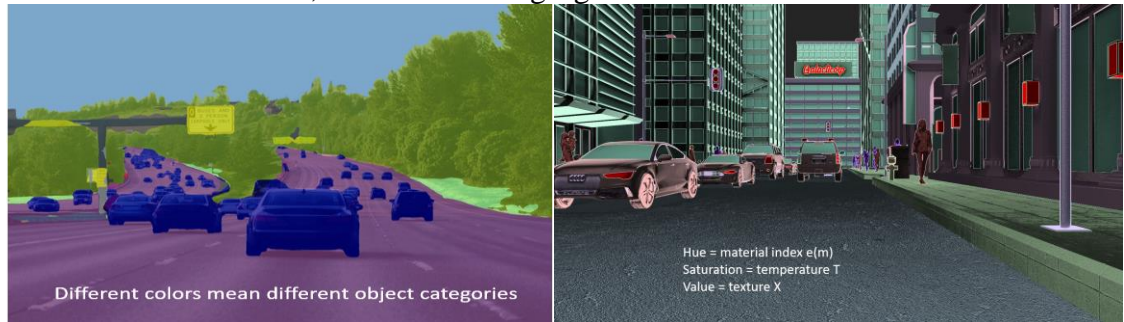


Figure 3. Left: a sample of the state-of-the-art semantic segmentation in the object level (ADE20K), with 6 categories. Right: a sample TeX vision in our HADAR database, **with 20 categories**. This comparison is to show the real-world complexity of our dataset.

The physical-component-level TeX vision will enable physics-driven semantic segmentation, as opposed to the state-of-the-art visual appearance driven semantic segmentation. We emphasize that TeX vision itself is not semantic segmentation and we have extensive algorithms to obtain the semantics of the scene from TeX vision, see Extended Data Fig.8 and Sec.SIIIE of the Supple. Info. for more details. In our city block dataset, there are multiple pedestrians, including men, women, kids, the elders, and robots, to mimic a future scene. We believe our database presents a convincing platform to test HADAR efficacy (TeX decomposition and TeX vision).

3. Based on our synthesized database, the machine learning performances of our TeX-Net (see Extended Data Fig.1) shows the applicability of TeX decomposition and TeX vision on complicated scenes. These results demonstrate HADAR efficacy on synthesized real-world scenes. The comparison of TeX-Net output with the ground truth TeX vision is given in Fig.S12 in the Supple. Info., as cited below.

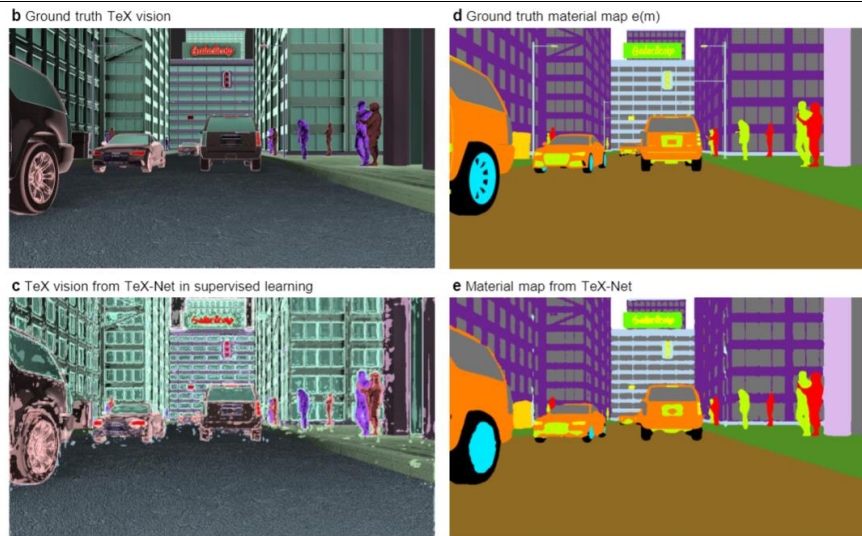


Fig.4 Comparisons of TeX-Net output with the ground truth show that TeX-Net is indeed able to perform TeX decomposition. TeX-Net exploits the spectral information in thermal radiation along with the HADAR constitutive equation (Eq. 1) to separate the intrinsic and extrinsic thermal photons. Small prediction errors in temperature lead to texture error in brightness, and hence there are some noisy spots observed in c. This can be improved by imposing sophisticated smooth constraint on temperature and harder training.

4. In this paper, we demonstrate a few typical examples, i.e., people detection with HOG+SVM, semantic segmentation with DANet, and ranging with DeepPruner. Performance comparisons between TeX vision and traditional thermal vision clearly indicates HADAR efficacy, as shown in Extended Data Figs. 5, 6, 8, and Fig. S17 of the Supple. Info. We briefly cite the results as below for the reviewer's convenience. We note that HADAR requires multi-spectral information to output TeX vision. **In the proposed concept of TeX vision, the scene is captured with physical attributes being represented by hue (emissivity index), saturation (temperature) and value (texture X).** This novel representation has physical context which is not present in the output of optical cameras (RGB vision), conventional IR thermal cameras (panchromatic thermal vision), or LiDAR (point cloud). Subsequent machine learning algorithms in computer vision regarding stereo matching, optical flow, scene flow, semantic segmentation, etc. that are previously based on RGB vision, thermal vision or point cloud can be adapted to TeX vision. Developing new algorithms exploiting TeX vision presents a new research frontier and we plan to pursue multiple avenues in future studies.

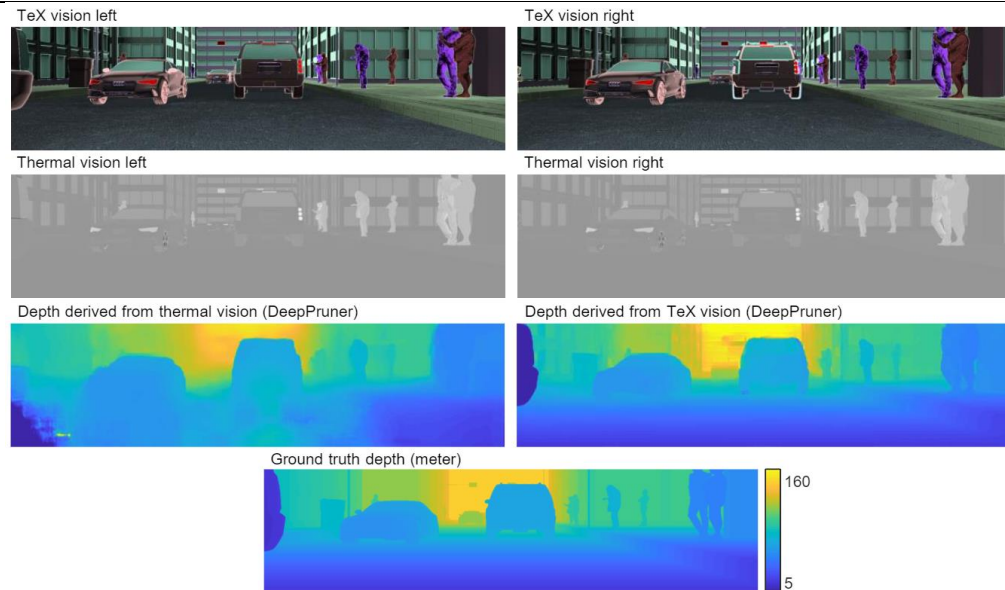


Fig.5 ‘TeX vision + AI’ beats the state-of-the-art ‘thermal vision + AI’ in ranging, showing HADAR efficacy in complicated scenes.

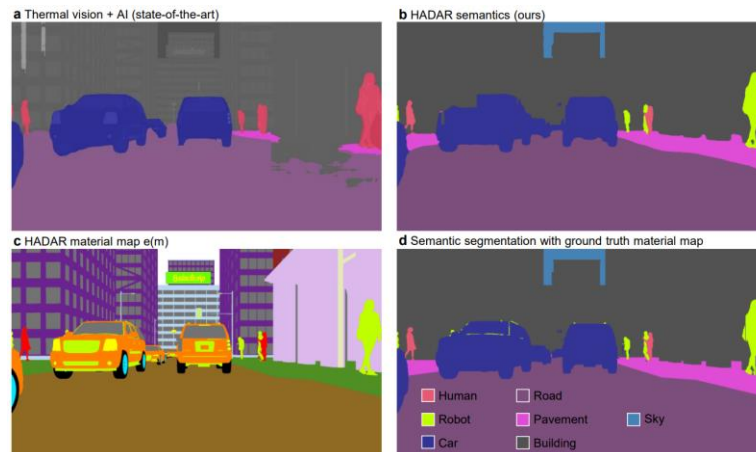


Fig.6 HADAR semantic segmentation based on TeX vision beats the state-of-the-art ‘thermal vision + AI’ segmentation, showing HADAR efficacy in complicated scenes.

5. We want to comment on the development of a real-world HADAR dataset. We note that a real-world experimental hyperspectral LWIR database with ground truth TeX vision and depth is not available. The procedures to build an experimental database are basically the same as what we have done in the outdoor experiments, and they can be briefly summarized as the following: (1) collecting left and right hyperspectral data cubes; (2) collecting the corresponding material library; (3) solving TeX as the ground truth by least-squares estimators; and (4) collecting ground truth depth with LiDAR. In this first paper, we have provided a synthetic database and shown two proof-of-concept experiments in summer and winter. This forms a foundation to develop a real-world HADAR dataset. The two key steps are outlined below.

Collecting material library: In our current experiments we used a subset of the NASA JPL ECOSTRESS spectral emissivity library as our material library. This library is for Spaceborne applications, not self-driving cars. Consequently, there are many materials (e.g., human skin, hair, clothes, and tires of cars) common in daily life but missing in the library, since they are rare to be seen from space. Instead, we have to use other similar materials to approximate the spectral emissivity. Note that TeX vision requires spectrally resolved emissivity different from existing panchromatic thermal vision where emissivity is approximated as a single number (i.e. $e(\lambda)$ vs $e=\text{constant}$). We did observe residual errors in our results due to the mismatch of emissivities used in the algorithm with respect to the actual emissivities (mismatch error remains in the texture map especially around boundaries in Fig.5 and Extended Data Fig.10; if the material library is perfectly known, one can recover texture as good as Extended Data Fig.2c). We intend to follow the same procedures as the JPL database [2] to generate a standard material database for self-driving applications. Handheld spectrometers can be used to collect the material library instead of bench-top spectrometers used in [2]. Building a material library includes spectrometer calibration, sample preparation, measurement, error analysis, and especially cross analysis with the JPL library for shared materials.

Spectral resolution of thermal image: Secondly, to distinguish more materials in the HADAR material library requires better spectral resolution, in our case, more spectral filters. However, research and commercialization of LWIR spectral filters currently lag behind visible-light filters. Especially in the COVID period, the 10 filters we used are the only significantly independent filters available in stock from Spectrogon, an industry-leading company providing LWIR filters. We are fabricating custom spectral metamaterial filters to enhance the resolution. Alternatively, using grating-/interferometer-based hyperspectral imagers could be another solution.

Our group has extensive preliminary work on those various aspects, and we are confident that the above two factors can be overcome in the near future. We do provide the research community with this first set of HADAR data collected in Indiana. However, building a standard material library or building a hyperspectral imager are independent projects beyond the scope of this first paper on HADAR.

Synthesized database is commonly used in the literature, for example, the Scene Flow dataset (<https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html>) for depth estimation and segmentation, and the MPI Sintel benchmark for optical flow (https://ps.is.tuebingen.mpg.de/research_projects/mpi-sintel-flow). Synthesized data, especially by Monte Carlo path tracing, has the real-world image quality and complexity, and has perfect ground truth and calibrations.

6. We also want to mention that our HADAR estimation theory is general for multiple scatterings and multiple objects, see Sec.SIII of the Supple. Info. Therefore, the HADAR efficacy for complex scenes illustrated above is consistent with the theory.

Reference(s):

	<p>[1] Fu, Jun, et al. "Dual attention network for scene segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.</p> <p>[2] A. Baldridge, et al, "The aster spectral library version 2.0", Remote Sens. Environ. 113, 711 (2009).</p>
C4	<p>For the development of HADAR, they have used FLIR A325sc, which is outdated. In addition, they have made the comparison with 16 channel lidar Velodyne puck, which gives sparse data. A comment about the functionality and cost comparison would help readers appreciate the proposed technique's significance.</p>
R4	<p>We thank the reviewer for pointing out the functionality-cost tradeoff. Indeed, reducing the cost while maintaining the essential functionality is crucial for our technology to achieve real-world impact. This is one of the guidelines we followed in designing those experiments.</p> <p>To develop HADAR, we need science-grade radiometric sensors as well as the spectral filters. The radiometric sensors that support quantitative analyses of the heat signal are preferred instead of 'thermal imagers' which focus on thermal visual appearances but distort heat information in the data.</p> <p>Camera Model: The FLIR Thermal Vision Automotive Development Kit (FLIR ADK, with Tau-2 camera core), for example, is a thermal imager that FLIR uses to generate the FLIR starter thermal dataset (https://www.flir.com/oem/adas/adas-dataset-form/). FLIR ADK is cost effective, around \$3,000, but it is less accurate and doesn't support radiometric analyses such as noise-equivalent temperature difference (NETD) [for radiometric analysis, ADK needs further calibration/upgrade which is more expensive]. In contrast, FLIR A325sc is the most cost-effective science-grade thermal camera available when we started the experiment one year ago. We note that now A325sc (320*240 pixel array, NETD<50mK, unit price around \$10,000) has been discontinued and covered by a more advanced model, A655sc (640*480 pixel array, NETD<30mK, unit price around \$20,000). But we emphasize that A325sc proves sufficient to demonstrate the prototype HADAR and illustrate its advantages.</p> <p>Spectral filters: The advantage of HADAR over traditional thermal vision comes from the spectral resolution and the theory we used to interpret the hyperspectral data. Therefore, we prioritized the optimization of filters, and chose almost all significantly independent filters from Spectrogon to fully fill the 12-position filter wheel (10 filters + Null + golden mirror; filters are around \$10,000 in total). Compact hyperspectral imagers are currently only used in defense applications as well as remote sensing and are very expensive. We believe our work will inspire the development of cheaper infrared spectral sensors for autonomous navigation applications. As another reviewer has also pointed out, the total cost of the HADAR system will greatly impact its applicability in real world. We think our current design using A325sc and spectral filters in our proof-of-concept experiments presents a good functionality-cost balance. With the development of relevant industry and new sensors, we anticipate the total cost would go down.</p> <p>Comparison to LiDAR: When we compare the prototype HADAR with LiDAR, our end goal is to show the advantage of HADAR in physics-based perception. Particularly, we want to show the evidence that (1) LiDAR can detect the cardboard and the human body but cannot distinguish them; and (2) LiDAR has a shorter detection range and has special difficulties in detecting low-reflectivity objects (e.g., very common black cars). Those two aspects are rooted in the operating principle of LiDAR (LiDAR measures geometric shape or depth through reflection signal) and do not depend on specific models. Therefore, we chose Velodyne Puck VLP-16 (\$4,000) which</p>

	<p>proves sufficient to compare with HADAR. Again, Velodyne HDL-32E (32 channels, \$15,000~\$20,000) or even Velodyne HDL-64E (64 channels, way more expensive) will surely give denser LiDAR data, but Puck VLP-16 presents a better functionality-cost balance. Another reason that led us to a sparse LiDAR is recent research developments called ‘pseudo-LiDAR++’ [1]. In pseudo-LiDAR++, researchers proposed to use sparse but cheaper LiDAR to de-bias stereo depth estimation and finally get dense and accurate ranging results. Pseudo-LiDAR++ inspired the possibility of combining sparse LiDAR with HADAR for some special applications. We are partially working along that line, but that is beyond the scope of this paper.</p> <p>As FLIR is constantly updating products in a pace that, unfortunately, customers cannot foresee, we have added a comment as suggested in the Method Section ‘Thermal camera specifications’ to clearly explain the datasheet of A325sc and the functionality-cost tradeoff, to help readers appreciate HADAR’s significance:</p> <p><i>“Our FLIR A325sc thermal camera is a science-grade high-performance radiometric camera (price ~ 10,000\$). It is equipped with an uncooled Vanadium Oxide (VoX) microbolometer detector that produces thermal images of 320 × 240 pixels. Detector pitch is 25 μm. Pixel size is approximately 12 μm. Time constant is 12 ms. Focal length is 18 mm. And f-number is 1.3. Noise equivalent temperature difference (NETD) is typically < 50 mK and characterized to be 47.8 mK. FLIR A325sc was available when the experiments in this paper were designed and conducted. We note that it has now been discontinued and replaced by a more advanced model, FLIR A655sc. The latter has a 640 × 480 pixel array with typical NETD < 30 mK, but it is twice as expensive. A better camera will give higher resolution HADAR data. Since the advantage of HADAR over traditional thermal vision comes from the spectral resolution and the theory we used to interpret the hyperspectral data, FLIR A325sc presents a better functionality-cost balance.”</i></p> <p>Reference(s): [1] You, Yurong, et al. "Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving." arXiv preprint arXiv:1906.06310 (2019).</p>
C5	<p>Authors state, “HADAR is distinct from hyperspectral imaging where material difference is determined by the Euclidean distance between their reflectance spectra [32]. In stark contrast, HADAR identifiability is determined by multi-parameter estimation of temperature, emissivity and texture”, but identifiability is estimated using a CNN with an input of proton profiles only (as given in the Methodology section). A precise statement would help readers to understand the implementation details.</p>
R5	<p>In the new version of our manuscript, we have expanded the Supple. Info. to include the full HADAR theory. We realize that it is better to understand the above statement with the help of the W-dimensional hyperspace (W being the number of spectral bands), see Sec.SIIA of the Supple. Info. For convenience, we explain the most relevant parts here.</p> <p>The difference in spectra between two materials in hyperspectral imaging (HSI) is described by the following paragraph (last paragraph, page 17 of the Supple. Info.):</p>

Start with hyperspectral imaging (HSI) where a library of reflectance spectra for potential materials, $\mathcal{R} = \{r_\nu^m | m = 1, 2\}$, is available. In the W -dimensional Hilbert hyperspace \mathbb{R}^W where each axis represents the reflection signal at certain wave number, $r(\nu_w), w = 1, 2, \dots, W$, the library \mathcal{R} is a set of two isolated dots, as shown in Fig. S5a. The difference between two materials is characterized by the Euclidean distance of those two dots [10, 11].

and the following figure of W -space:

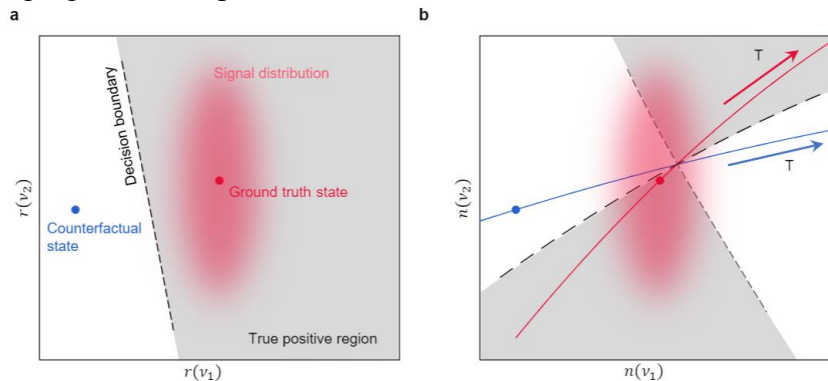


Figure 7. Part of Fig.S5 of the Supple. Info. (a) material difference of HSI is the Euclidean distance between red and blue dots. (b) material difference in HADAR is the shortest ‘distance’ of the red dot to the blue curve/sheet.

The identifiability between two materials in HADAR is described by the following two paragraphs (second paragraph, page 19; line 4, page 25):

Because temperature T and thermal lighting factor V are both unknown in HADAR, one cannot directly get spectral emissivity curve out of $S_{\alpha\nu}$ in Eq. (S2). Instead, we have to move to the W -dimensional Hilbert hyperspace \mathbb{R}^W where each axis represents the heat signal $n(\nu_w), w = 1, 2, \dots, W$, as shown in Fig. S5b. Here, materials at fixed V and T are isolated dots, but each material corresponds to one sheet of V - T surface given by Eq. (S18), and generally different material surfaces might intersect.

Since temperature T and thermal lighting factor V are both unknown, HADAR has to estimate multiple parameters $\{mTV\}$ simultaneously from heat signal S to identify the material. The material difference at the given target state (red dot) is no longer characterized by the Euclidean distance between the red dot and the blue dot (with the same V and T as the red dot). Instead, it is characterized by the shortest ‘distance’ of the red dot to the blue curve, where the shortest distance is exactly captured by the multi-parameter statistical (Mahalanobis) distance.

	<p>TeX-Net takes hyperspectral data cube as input and performs multi-parameter estimation of three physical attributes for each pixel of the scene – T, e and X. For non-neural-network approaches of TeX decomposition working on each pixel, we take the input as the radiation spectrum S and the output is TeX values, (see Methods --- TeX decomposition), of which $e(m)$ describes the material category. The three physical attributes are coupled and hence uncertainty in texture and temperature will directly affect the accuracy of material classification (see extended data figure 13).</p> <p>To make it easier for readers to understand the argument, we have revised the statement in the main text to guide the readers to the W-space and relevant contexts in Supple. Info.:</p> <p>“HADAR is distinct from hyperspectral imaging where material difference is determined by the Euclidean distance between their reflectance spectra. In stark contrast, HADAR identifiability is determined by multi-parameter estimation of temperature, emissivity and texture (Fig.S5 and relevant contexts in Supple. Info.)”</p>
C6	<p>Authors claim, “The minimum photon number for given semantic distance or vice versa, the minimum semantic distance for given photon number sets fundamental limits to object identification beyond training volume, providing a theoretical foundation for designing public policies.” However, ML algorithms generally perform well with missing or scarce information; placing a bound on input data is yet an open challenge. A comment about the significance of bounds would help the cross-discipline reader base.</p>
R6	<p>We agree with the reviewer that machine learning algorithms generally perform well with missing or scarce information, such as, for image generation, data prediction, data completion, denoising, etc. However, our perspective is different from the above. Our purpose is exactly to place a bound on the ML performances (on average) determined by the input data quality.</p> <ol style="list-style-type: none"> 1. Shot noise limit of machine learning: As the input data is usually an optical signal, we would like to further explain our objective with a concrete example of daily experience. Suppose we have a phone camera taking a picture of a barcode and our task is to identify the barcode to retrieve the associated information, see the following figure. This scenario is common in a supermarket or a library, and the barcode is actually a spatial profile analogue of the spectral profiles under consideration in our manuscript. General machine learning usually uses sufficiently strong signal (photon number is huge), and it is good at dealing with blur, missing, or noisy signal (see the following Fig. a-c). However, one always wants the identification process to be as fast as possible to be more productive (by decreasing the exposure time), and the illumination light to be as weak as possible to save energy, and even the camera aperture to be as small as possible to make the phone more compact. Pushing toward those limits, the signal becomes so weak that the light field should be treated as quantized photon streams. In this limit, the input data will become poorer and poorer (d-e) and eventually fail the identification test. <u>What are the criteria to which we can optimize the hardware design (phone) while still guaranteeing successful identification?</u> Our manuscript is devoted to addressing this type of question, in the context of HADAR, with all parameters modelled in the total photon number.

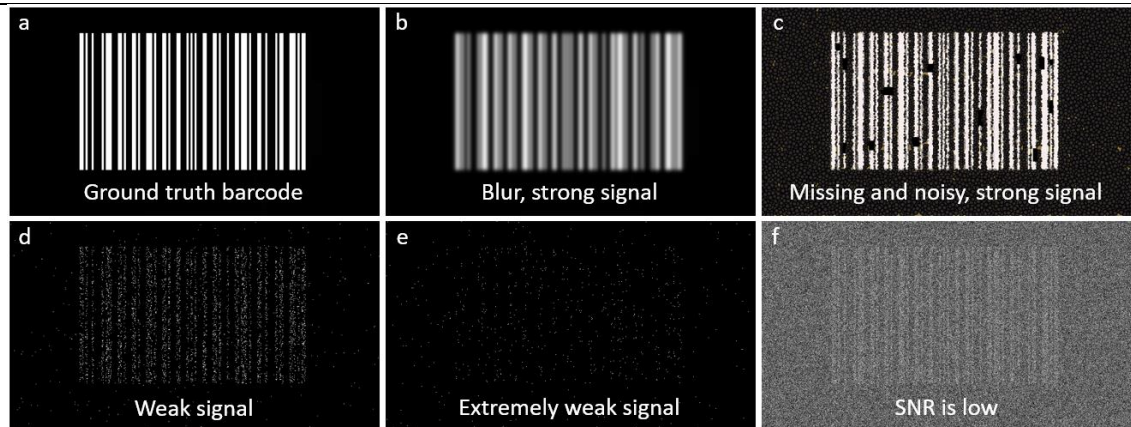
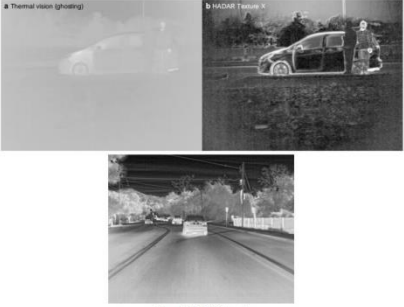


Fig.8 General machine learning performs well on blur, missing, noisy, but strong signal (a-c). When the input signal is extremely weak (a few photons, e) or the signal-to-noise ratio (SNR) is extremely low (f), the quality of input data is so poor to train a neural network well, no matter how large the training volume is. This figure is generated particularly for the reviewer, not included in the paper.

Few-Photon-Limit: The input data for training vision algorithms consists of images. These images are bitstream of photons falling on a detector. In the quantum limit, only a few photons fall on a detector. Our test data and training data are in this few-photon-image limit. In this limit, neural network performance is bounded (on average over many runs) by the fundamental physical laws of thermal photonic information theory. Thus, we can use the bound to compare across different neural network architectures and whether they can indeed achieve the bound in tasks like object detection/classification/ranging etc.

Detector noise: The bound specifically includes the detector specifications used to generate the training data. The inefficiencies of the detector used to collect training data causes a limitation in the best performance achievable by the neural network independent of the volume of training data.

2. In comparison, missing data in general machine learning is usually caused by detector readout failure, post-processing bugs, human errors, etc. But our focus is on the total information input to the detector determined by the hardware configuration. Detector imperfection (readout failure or electronic noise) is included in our theory, while human error or software bugs are not. Human error or software bugs are practical factors that will degrade the data quality, but our fundamental bounds are useful because they depict the optimal performances of machine learning when human error and software bugs are completely corrected. The **shot-noise limit** further depicts the optimal performance of machine learning when the sensor noise is optimized. Practical errors show variability across systems, but our shot-noise limits which exploit physical laws of thermal information theory are universal and can be the guidance to public policies.
3. **Cramer-Rao Bound for Machine Learning:** Any machine learning algorithm or its associated neural network is a particular estimator to the problem under consideration. The Cramer-Rao lower bound is the lower bound to the uncertainty of any unbiased estimator. We agree that placing a bound on one single ML evaluation is an open

	<p>challenge to date, but on average, ML performances will be bounded by the Cramer-Rao bound, as we have demonstrated in Fig.3b in the main text. Therefore, we think our theory is important as it combines machine learning with information/estimation theory, and it lays a theoretical foundation (1) for analyzing the maximum achievable machine learning performance given a specific detector and (2) for designing public policies of minimum hardware requirements for a given ML task.</p> <p>To make our argument clearer, we have added a comment in Methods --- Guiding public policy (2nd paragraph):</p> <p>“... <i>Our fundamental limits bound the average machine learning performances due to the shot noise and detector noise. Specific lucky evaluation events can occur but they will fluctuate around the average bounds, as can be seen in Fig.3b and insets of Fig.4. Human error or software bugs are not considered in our bounds, but our bounds are useful because they depict the optimal performances of machine learning when human error and software bugs are completely corrected. Therefore, our bounds related to physical laws of thermal photonic information theory can be used as a guidance to public policies.</i>”</p>
<p>C7.1</p>	<p>Authors said that “Thermal imaging loses textures due to TeX degeneracy (Fig. 4a) and leads to inaccurate ranging”. Thermal cameras, for instance, FLIR BlackFly (BFS-U3- 51S5C-C), produce impressive images with texture, as shown in figure-1.</p>  <p>From FLIR Dataset Figure-1</p>
<p>R7.1</p>	<p>We agree with the reviewer that in the FLIR dataset (panchromatic thermal vision) there are impressive thermal images with textures. We would like to clarify our argument about texture by (1) explaining the relation of our work of TeX vision with the FLIR approach; (2) explaining how the FLIR approach works; and (3) making a fair comparison of our work with the state-of-the-art FLIR approach. Furthermore, we have also obtained new experimental results with improved texture recovery.</p> <ol style="list-style-type: none"> 1. TeX vision exploits the spectral information in thermal radiation along with the HADAR constitutive equation (Eq. 1) to separate the intrinsic and extrinsic thermal photons. This is in stark contrast to the FLIR dataset which is panchromatic i.e. all spectral information is lost at the detector itself. In TeX vision, the scene is captured with physical attributes being represented by hue (material index), saturation (temperature) and value (texture). This new representation has physical context which is not present in

the output of optical cameras (RGB vision), conventional IR thermal cameras (panchromatic thermal vision), or LiDAR (point cloud).

As shown in the texture flow diagram below, thermal textures in the scene will be (1) collected by hardware sensors and (2) visualized to users. In Sec.SID of the Supple. Info., we theoretically analyzed the heat signal and identified 3 types of thermal textures in the scene, i.e., T-type (temperature contrast), e-type (nonuniform material), and X-type (geometric texture). Moreover, we listed 4 texture-loss channels in data collection, i.e., shot-noise/detector noise, finite bit depth, spectral integral, and the TeX degeneracy. Thermal imaging loses textures in step one, data collection, but can still manifest certain residual textures in practice (FLIR dataset). The amount of residual texture is scene dependent. For special scenes, such as the HADAR alphabet sample shown in Fig.6b or the car & pedestrian scene in Fig.4a in the main text, this residual texture may completely vanish. In general, even the collected residual texture in thermal data cannot be seen by users if step two, data visualization, is improper.

We note that BlackFly BFS-U3-51S5C-C is an optical camera in the visible-light spectrum, not a thermal camera, so we focus our response on FLIR ADK (with Tau-2 camera core) that is used to generate the FLIR dataset. It uses a sophisticated AGC (automatic gain control) algorithm to improve the visual contrast in step two, maximizing the usage of residual texture. We emphasize that image processing like AGC cannot add textures to the data but can only maximize the visualization of existing textures. This is implied by the well-known Data Processing Inequality --- 'post-processing cannot increase information' [1]. In comparison, HADAR is improving step one, data collection, by estimating the spectral dependence of blackbody radiation law as well as the spectrally resolved reflectivity and emissivity of materials. This is crucial to increasing the information content (texture) in the collected signal which is otherwise lost in traditional panchromatic infrared imaging.

In the visualization phase, we use AGC as well (see Sec.SIIC of Supple. Info.) to maximize the visual effect of collected textures. Various artificial intelligence algorithms may have various performances in detection and ranging, but the fundamental bound is determined by the raw sensor data input to neural networks. Therefore, HADAR which improves data collection through spectral resolution has better fundamental information content about the heat signal than thermal imaging. Our theory of thermal textures and the following figure have been detailed in Sec.SID of the Supple. Info.

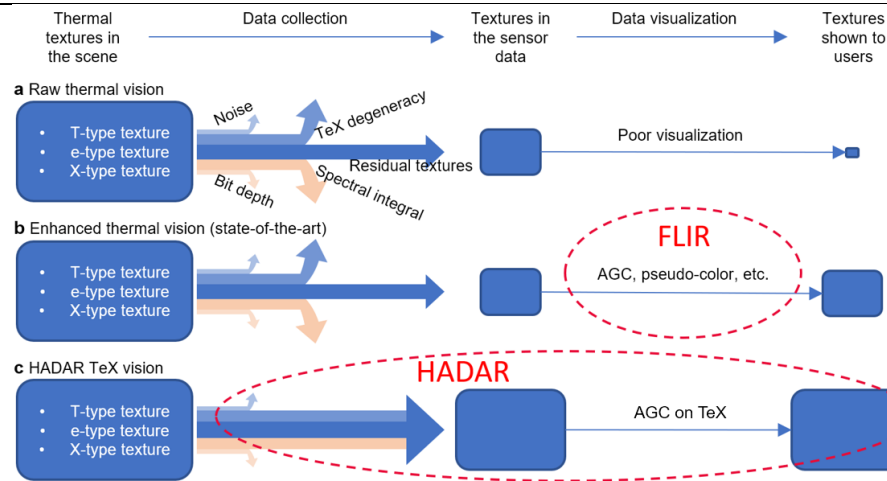


Fig.9 Texture flow diagram to compare our work with the FLIR approach. HADAR is collecting more information in the sensor data, in contrast to state-of-the-art approaches of post image processing.

- The impressive image quoted by the reviewer from the FLIR dataset (FLIR ADAS 1_3/FLIR_00316) has a low-contrast raw vision before image processing, see Fig.10a below. The raw data output by FLIR camera is also provided in the FLIR dataset in the 16-bit raw data folder. Raw thermal data is distributed in two narrow pixel value ranges, as can be seen in the histogram ©. Details with small pixel-value variations are difficult for human eyes to perceive but can be discriminated by machines. FLIR AGC attempts to enhance the visual contrast of weak data variations. AGC is a modified version of histogram equalization algorithm (using high-pass filters before histogram equalization). It performs a nonlinear map from raw pixel values to new pixel values so that the new image has a relatively uniform distribution in histogram, see Fig.10d. AGC is not a unique mapping but has multiple free parameters, see the right panel of Fig.10. FLIR also states in the FLIR Application Notes (<https://flir.netx.net/file/asset/15755/original>), ‘FLIR highly recommends that each customer optimize AGC parameter settings for each particular application. “Preferred” AGC settings are highly subjective and vary considerably depending upon scene content and user preferences’. From the above details, it can be seen that AGC is a heuristic post-processing algorithm that improves visual contrast but cannot increase information in the sensor data. On the other hand, TeX vision is a representation of infrared heat radiation where hue-saturation-value in the image is **thermal-physics-driven**.

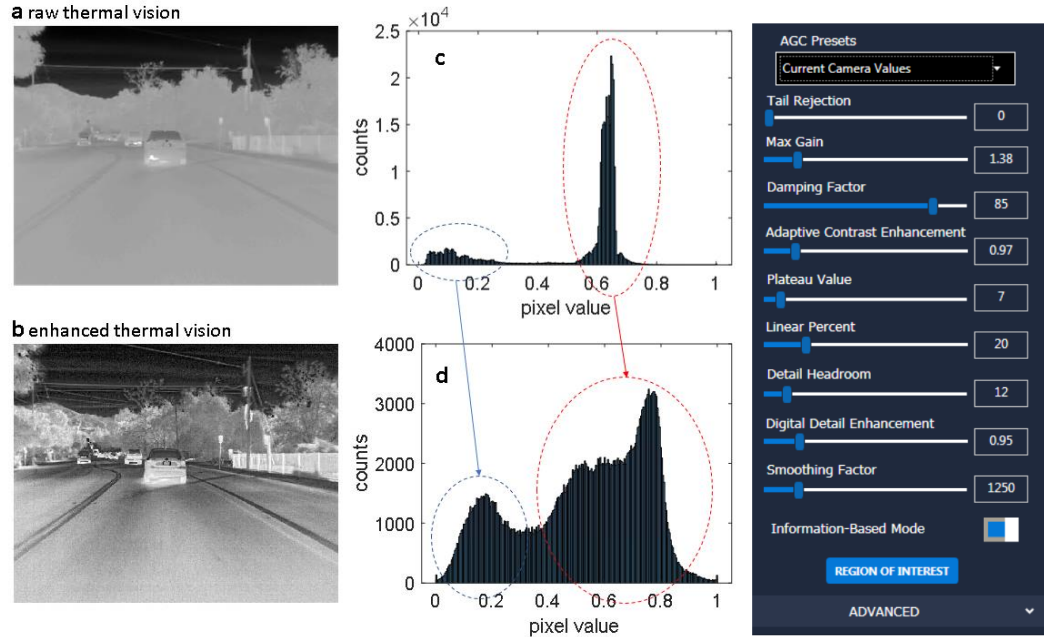


Fig.10 (a) raw thermal vision from the FLIR dataset. (b) Enhanced thermal vision with AGC, as quoted by the reviewer. (c) histogram of (a). (d) histogram of (b). The right panel is the FLIR AGC template, showing multiple parameters to improve visual contrast

3. To make a fair comparison of our work with the state-of-the-art approach, we will quantify textures at both the fundamental level and the visual level. In Sec.SIID of the Supple. Info., texture quantification, we have introduced two metrics. The Fisher information metric quantifies textures at the fundamental level and relates to the ranging accuracy, while the standard deviation metric quantifies textures at the visual level. At the fundamental level, we derived the expression of texture, J_0 , in Tab.4 of Supple. Info. with (HADAR) or without (thermal imaging) spectral resolution. The fact that textures with spectral resolution are more than textures without spectral resolution can be rigorously proved and understood, as explained by the paragraph below Eq.S43 of the Supple. Info.

Eq. (S43) recovers Rayleigh's limit. We now briefly prove that the Fisher information for HADAR ranging with spectral resolution is more than Fisher information for panchromatic thermal imaging. In the mathematical expression of the Fisher information for HADAR in Tab. S4, the spectral information is squared before integral, which prevents destruction of the spectrally resolved Fisher information from contributions of opposite signs. This leads to a larger Fisher information J_x^0 and a smaller photonic correspondence uncertainty σ_c . Mathematically, $\int \frac{(\partial_x p_{x\nu})^2}{p_{x\nu}} d\nu - \frac{(\int \partial_x p_{x\nu} d\nu)^2}{\int p_{x\nu} d\nu}$ can be manipulated into a square form, $(*)^2 \geq 0$, $*$ being a certain expression, and hence it proves that the Fisher information is larger with spectral resolution. More importantly, by breaking the TeX degeneracy, HADAR can support sophisticated priors like sparsity or smoothness to further remove unknowns in the parameter set $\{m_\alpha, T_\alpha, V_\alpha\}$, suppressing ranging error toward a lower bound, $J_x^0 \leq \iint \Omega \frac{(\partial_x b_{x\nu})^2}{b_{x\nu}} + \frac{(\partial_x k_{x\nu})^2}{k_{x\nu}} ds d\nu$, with $b_{x\nu} \equiv \tilde{S}_{x\nu}^0 / \iint \Omega S_{x\nu} ds d\nu$ and $k_{x\nu} = p_{x\nu} - b_{x\nu}$. Here, $\tilde{S}_{x\nu}^0$ is the direct emission.

We note that the Fisher information in the heat signal is defined on the raw sensor data. This fundamental metric is not altered by post-processing techniques like AGC and can only be governed by the detector physics. This Fisher information metric governs the error in the accurate estimation of texture. Therefore, AGC-enhanced thermal vision cannot increase Fisher information as AGC does not alter the sensor itself. It has less Fisher information than HADAR since our approach alters the sensor and adds spectral information. Thus the Fisher information metric shows that HADAR recovers more textures than thermal imaging.

To make a fair comparison between the FLIR dataset and our TeX vision approach at the visual level, we first explain the concept of pseudo-coloring. We use the pseudo-coloring approach as the state-of-the-art reference instead of the grayscale AGC approach since our TeX vision has 3 color channels. We use these 3 color channels for both our work and the state-of-the-art. We note that pseudo-coloring has the same functionality as AGC, and pseudo-coloring is more commonly used in FLIR camera visualizations. AGC is used to map pixel values from grayscale to grayscale, but pseudo-coloring is utilized to map pixel values from grayscale to RGB triplets. Both of them can enhance visual contrasts, see Fig.11 and also the FLIR Application Notes (<https://flir.netx.net/file/asset/15755/original>).



Fig.11 State-of-the-art approaches to improve visual contrast. To visualize more details, (b) AGC maps raw pixel values from grayscale to grayscale, while (c) pseudo-coloring maps raw pixel values from grayscale to RGB.

For the outdoor experiment in Indiana in winter, we quantified textures with the standard deviation metric (stdfilt function in matlab) for our TeX vision (Fig.S11 d-f of the Supple. Info., as cited below), compared to the state-of-the-art (a-c). We can see that TeX vision (0.077) almost has 3 times more textures than the state-of-the-art enhanced thermal vision (0.027).

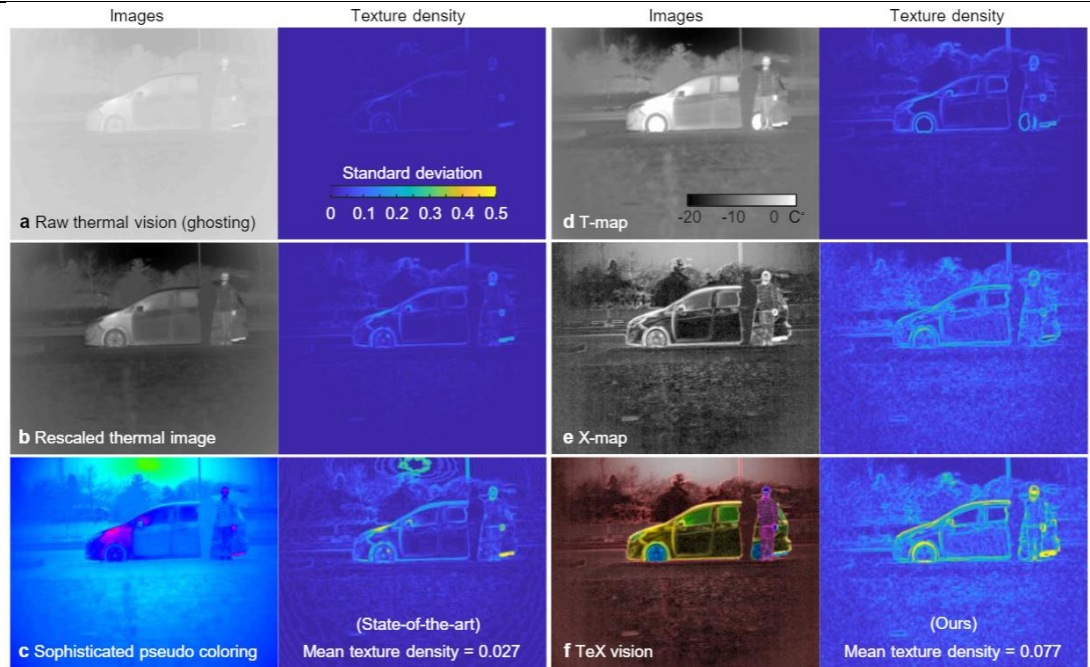


Fig.12 Quantitative comparison of our TeX vision with state-of-the-art thermal vision regarding texture recovery. HADAR TeX vision almost has 3 times more textures than the state-of-the-art enhanced thermal vision. Experiment was done at a winter night (Dec. 2020).

For special scenes like the car & pedestrian scene in Fig.4 of the main text, once textures are lost in the sensor data, neither AGC nor pseudo-coloring can recover the texture, see below

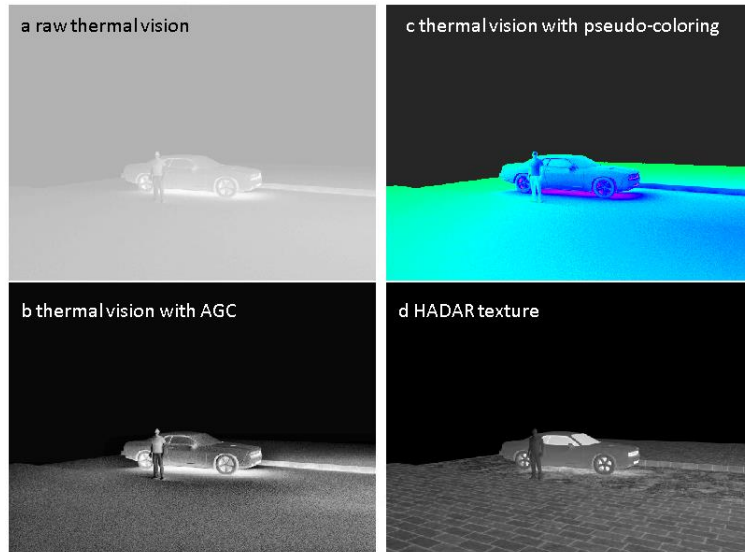


Fig.13 Comparison of HADAR texture with state-of-the-art approaches in recovering textures. This figure is given as Extended Data Fig.3. AGC is post processing. Once texture is lost in sensor data, AGC cannot recover the texture. In comparison, HADAR has spectral resolution and collects more textures in sensor data itself, enabling recovery of otherwise inaccessible textures.

Therefore, the argument that HADAR recovers more textures than thermal imaging with AGC also holds at the visual level.

4. To improve our results about texture recovery, (1) we have done one more experiment in summer daylight, and (2) we tested texture recovery on our HADAR city block database.
 - The summer daylight experiment below shows that HADAR (0.0879) recovers more textures (4.6 folds) than the state-of-the-art pseudo-color approach (0.0191).

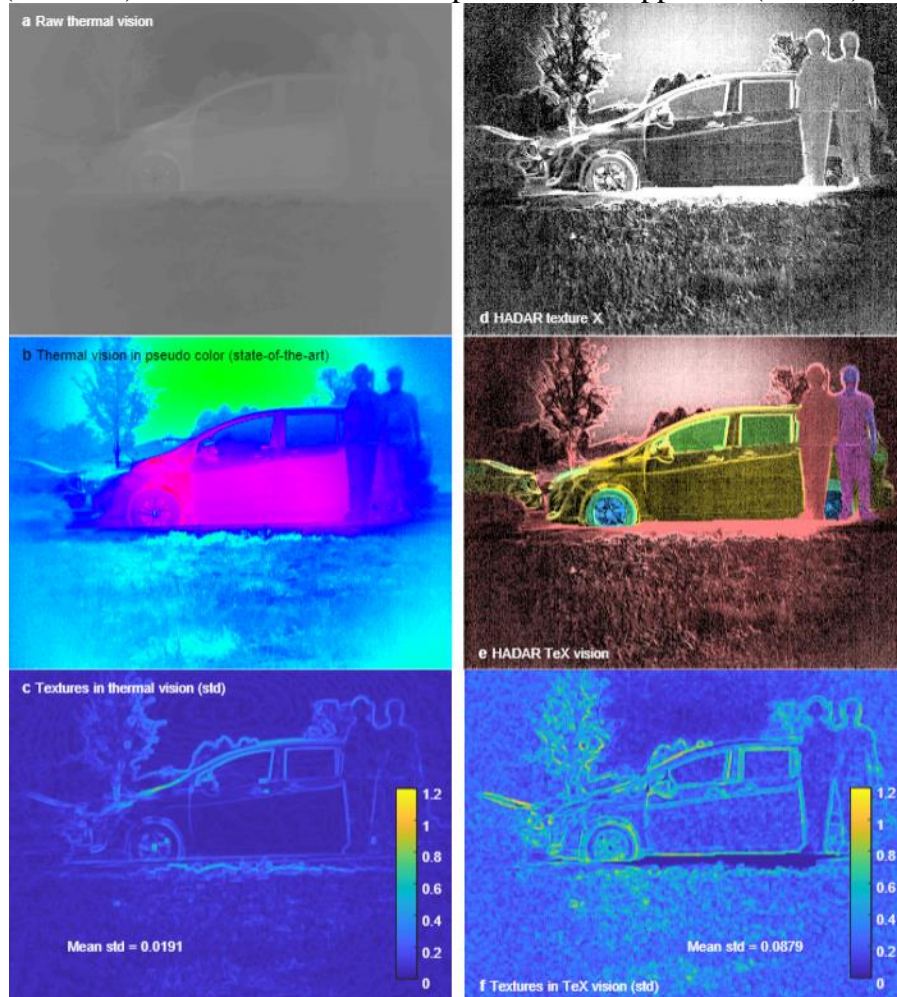


Fig.14 HADAR TeX vision in summer daylight. HADAR TeX vision has 4.6 times more textures than the state-of-the-art enhanced thermal vision. This figure is a new experiment result given in the Extended Data Fig.10.

- We tested our TeX vision on the city block dataset, as shown below. The city block dataset clearly shows that TeX vision recovers textures, beating both raw thermal vision and state-of-the-art enhanced thermal vision. Quantitatively, the mean texture density in enhanced thermal vision (standard deviation metric) is 0.0170, while the mean texture density in TeX vision is 0.0788 and is about 4.6 folds larger. This result is given in Extended Data Fig.2. In the dataset, emissivity in material library is

accurately known. Also, image size of the dataset is 1080*1920, much larger than FLIR A325sc. These two factors make the TeX vision in the synthesized dataset much better than proof-of-concept experimental performance of TeX vision.

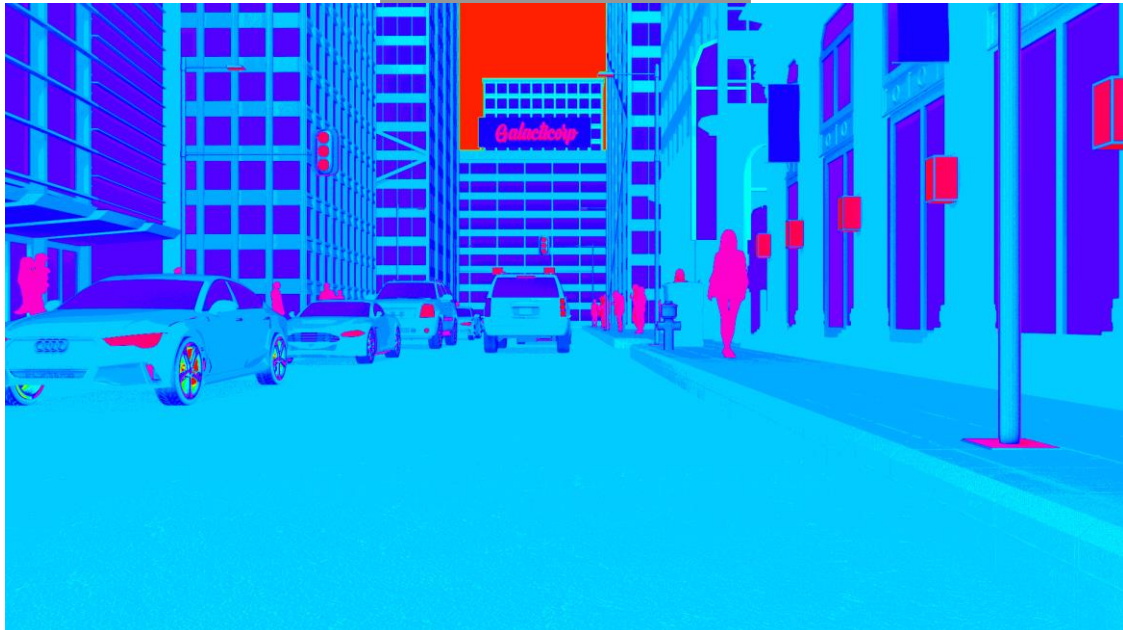


Fig.15 Top: Raw thermal vision with ghosting effect. Middle: State-of-the-art enhanced thermal vision in pseudo color. Bottom: HADAR TeX vision. HADAR TeX vision has 4.6 times more textures than the state-of-the-art enhanced thermal vision. To better visualize textures, we enlarge TeX vision in comparison with state-of-the-art thermal vision. Texture density figures are not shown here.

- Furthermore, we would like to add a comment about FLIR that (1) the thermal imager used in the FLIR dataset, FLIR ADK, has a larger pixel array than our FLIR A325sc (640*480 vs. ours 320*240). This can usually make images more impressive, as discussed above with our dataset. Based on the same camera and same condition, HADAR with spectral resolution will be better in texture recovery than thermal imaging without spectral resolution. (2) The key to improving visual contrast is to subtract the strong signal floor and keep weak variations. FLIR AGC and pseudo-coloring are empirical approaches to subtract the signal floor, as stated in the above FLIR Application Notes. In comparison, HADAR measures temperature and emissivity to estimate the direct emission which is exactly the strong signal floor. Hence, HADAR is a physics-inspired way to subtract the signal floor, see more discussions in Sec.SIIC of the Supple. Info.

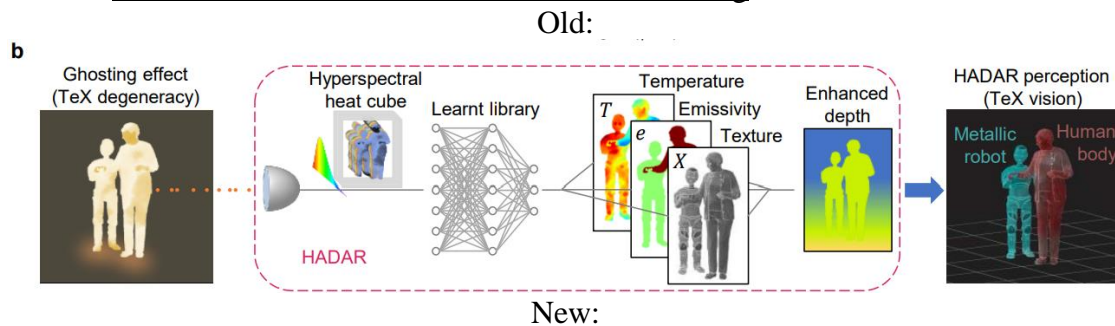
Reference(s):

[1] N. J. Beaudry and R. Renner, An intuitive proof of the data processing inequality, Quantum Information & Computation 12, 432 (2012)

C7.2 Moreover, in this work, emissivity is used for range computation instead of texture.

R7.2 We note that we use all three coupled physical attributes of TeX vision: temperature, emissivity, and texture for computing the range. TeX vision exploits spectral information in infrared heat radiation along with the HADAR constitutive equation (Eq 1) to separate intrinsic and extrinsic thermal photons. This thermal-physics-driven approach gives rise to three information channels which we represent as hue (emissivity), saturation (temperature) and value (texture). This is fundamentally different from optical cameras which output RGB vision. HADAR ranging is based on stereo matching of left and right TeX vision images, like traditional stereo matching on RGB visions. Thus, the large number of neural network architectures used in optical vision tasks can be adopted to TeX vision in the near future.

(1). We apologize that in the previous version of Fig.1b , it was confusing to put depth before TeX vision. In this new version, we revised it to the following.



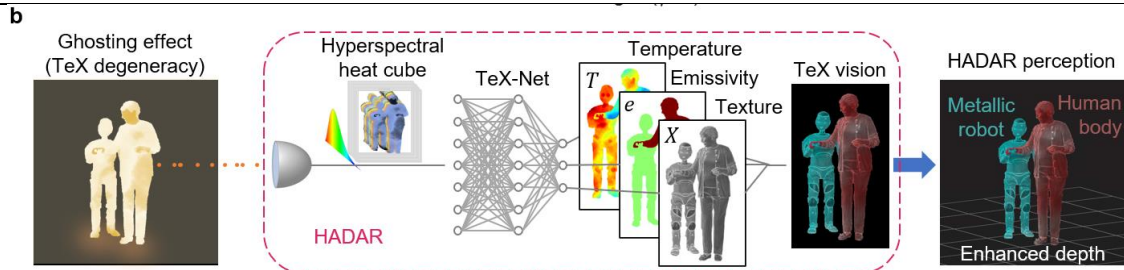


Fig.16 Top: previous Fig.1b. Bottom: new Fig.1b. HADAR generates TeX vision. Detection, semantic segmentation, ranging, etc are based on TeX vision. Depth is computed after TeX vision.

(2). To make our approach clearer, we have directly stated in the introduction (last sentence, 1st page) that:

“Our demonstrations of HADAR includes detection and ranging based on TeX vision, for both real-world level HADAR database and outdoor experiments.”

(3). In HADAR ranging section of the main text and Fig.4, stereo matching is used only for the scattering signal, $(1-e)X$, to show the importance of texture in ranging and compare with optical imaging. This is because the main text focuses on describing the fundamental limits of HADAR, and it is more intuitive to illustrate the ranging error bound with only the scattering signal. However, we emphasize that our fundamental bound on ranging error is universal and applies to all kinds of images, including the TeX vision, X-type texture, the scattering signal, or even optical images. We have added these explanations in the HADAR ranging section to make the logic flow more fluent:

“...To show the importance of texture in ranging and compare with optical imaging, here we focus on the scattering signal that can be reconstructed through TeX decomposition...”

(4). Moreover, HADAR ranging based on TeX vision is also explicitly demonstrated in Extended Data Fig.6. Stereo matching is performed with DeepPruner pre-trained on the KITTI dataset.

C8	An example where thermal equilibrium can cause singularity would help appreciate the utility of identifiability and ranging.
R8	<u>We have added two typical examples of equilibrium singularity</u> in Sec.SIIA, page 25 of the Supple.Info., where we discussed the HADAR identifiability in details, to help readers appreciate the utility of identifiability. The relevant contents are cited below. In thermal equilibrium, singularity means it is impossible to distinguish objects’ materials.

	<p>the target ($P \rightarrow 1$) through measuring the frequency. In the opposite limit where thermal lighting factor $V_\alpha = 1$ and $T_\alpha = T_0$, one can show semantic distance $d_0 \rightarrow 0$, and hence it is impossible to identify the target ($P \rightarrow 1/2$) no matter how many photons we have. We call the latter situation an equilibrium singularity. In this case, any target would form a cavity and be in thermal equilibrium with the environment, with photon number of the radiation field given by Boltzmann's distribution. It consistently leads to the blackbody radiation spectrum $S_{\alpha\nu} \equiv B_\nu(T_\alpha)$ as given in Eq. (S18), and hides every material feature of the target. One typical example of the equilibrium singularity is the standard cavity-based blackbody source commercially available. As long as the cavity is enclosed (with a tiny hole, $V_\alpha \approx 1$, $T_\alpha = T_0$), the output spectrum is a blackbody spectrum whatever the material is used inside the cavity. Another phenomenon of the equilibrium singularity can be commonly seen in a closed office room ($V_\alpha = 1$, $T_\alpha \approx T_0$). Thermal imaging of walls, desks and chairs inside the office (observed, not shown) would appear uniform of no texture even they are made of very different materials. The objects are indeed emitting different amounts of thermal radiation. However, the scattered signal from the environment which is in thermal equilibrium with the object completely balances these differences. Numeric</p>
C9	<p>A comment about the change in-bounds considering the non-stationary objects in a scene would help assess TeX utility in interpreting sequential information.</p>
R9	<p>We thank the reviewer for asking about the TeX utility for non-stationary objects and sequential information. Indeed, this is the common case in applications when either the object or the intelligent agent equipped with HADAR is moving. The relative motion of objects is described by scene flow in the literature. Projected onto the image plane, scene flow is manifested as motion blur in each individual image and optical flow in sequential image frames. Now, we discuss the bounds of HADAR detection and ranging, TeX decomposition and TeX vision, in the presence of scene flow with different motion-blur levels.</p> <ol style="list-style-type: none"> 1. Weak motion blur: Although our bounds and TeX vision are derived and demonstrated for stationary objects, they are also applicable for non-stationary objects when the motion blur is negligible, that is, when the apparent motion of a point source is within one pixel on the image plane. The apparent motion is given by $\Delta = vtL/r\theta$, where v is the relative transverse speed, t is the exposure time, L is the number of pixels in the horizontal direction, r is the distance of the target, and θ is the field of view. Motion blur is negligible when either the transverse speed is low or the exposure time is short. For example, a target at 30 m away captured by FLIR A325sc ($t < 12$ ms, $L = 320$, $\theta = 50$ degree) equipped on a car driving at 30 mph [$v \leq 30 \sin \frac{\theta}{2}$ mph] will have $\Delta \lesssim 0.8$ and hence the motion blur is negligible. To allow a higher travelling speed, the hyperspectral data cube acquisition rate of the used camera must be high so that the exposure time is sufficiently short to avoid motion blur, according to the criterion $\Delta < 1$. This criterion, $\Delta < 1$, constrains the applicability of our bounds. Within the criterion, TeX decomposition can be performed for each individual heat cube to obtain TeX vision. Subsequent detection and ranging are based on TeX vision. Worth noting is that traditional optical flow, scene

	<p>flow, semantic segmentation, etc., can all be extensively explored based on TeX vision and depth, presenting a new research frontier. For example, the RGB-d flow in Ref. [1] can be formally transplanted on TeX vision and depth (TeX-d), to retrieve sequential information.</p> <ol style="list-style-type: none"> 2. Moderate motion blur: For stronger motion blur beyond the criterion, if local motion field can be represented by linear convolutional kernels, there are multiple motion-blur removal algorithms available to estimate the motion field [2,3,4] and get the clean signal without motion blur out of the raw data. Consequently, TeX decomposition and TeX vision are applicable again after the pre-processing of motion-blur removal. 3. Strong motion blur: In the limit of extremely long exposure time, the motion blur kernel is a complicated convolution depending on the velocity field of the scene flow. The algorithms to remove motion blur in the presence of such motion blur are still open questions and deserve future research. New generation of ultrafast bolometers and hyperspectral imagers can mitigate strong motion blur for moderate navigation speeds. 4. However, in the presence of strong motion blur, the bound for ranging accuracy (Eq. S43 of Supple. Info.) still holds, even though the photonic correspondence uncertainty now includes contributions from motion blur in a complicated form. In this scenario, we can directly use Eq. S37, which is universal for all stereo images (including those with motion blur) and can be derived for given image pairs themselves. <p>Accordingly, <u>we have added the above comment in Sec.SIIE of the Supple. Info. --- bounds in the presence of scene flow.</u></p> <p>Reference(s): [1] Herbst, Evan, Xiaofeng Ren, and Dieter Fox. "Rgb-d flow: Dense 3-d motion estimation using color and depth." 2013 IEEE international conference on robotics and automation. IEEE, 2013. [2] Sun, Jian, et al. "Learning a convolutional neural network for non-uniform motion blur removal." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. [3] Gupta, Kavya, Brojeshwar Bhowmick, and Angshul Majumdar. "Motion blur removal via coupled autoencoder." 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017. [4] Portz, Travis, Li Zhang, and Hongrui Jiang. "Optical flow in the presence of spatially-varying motion blur." 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012.</p>
C10	<p>D. Appropriate use of statistics and treatment of uncertainties</p> <ol style="list-style-type: none"> 1. They have claimed that they have achieved 100 x accuracy in HADAR ranging, which is physics-based semantic segmentation between a person and a metallic body. In regard to computer vision literature, AI-based semantic segmentation results are already established. They have not made a comparative analysis between their proposed method and state-of-the-art AI-based semantic segmentation. Second, they have done the semantic segmentation using emissivity, and if the two subjects have the same emissivity, then their method fails. Below are some references for the AI- based semantic segmentation on thermal images (1) Li, Chenglong, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation." IEEE Transactions on Neural Networks and Learning Systems (2020).

(2) He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.

(3) Treible, Wayne, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu."Cats: A color and thermal stereo benchmark." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2961-2969. 2017.

R10 We thank the reviewer for drawing our attention to the above references and we agree that it is necessary to make a fair comparison to the established AI-enhanced thermal semantic segmentation. We checked the above references trying to get a proper baseline for comparison. However, the authors of the above Ref. (1) replied that their model is not saved and unavailable. The above Ref. (2) is for instance segmentation and the above Ref. (3) is for stereo matching. Instead, we turn to the DANet [1] which is also one of the state-of-the-art AI-based semantic segmentation that can be applied on thermal images.

- The quantitative comparison is added in Extended Data Fig. 8, as cited below. It can be clearly seen that our physics-driven semantic segmentation gives better segmentation results.

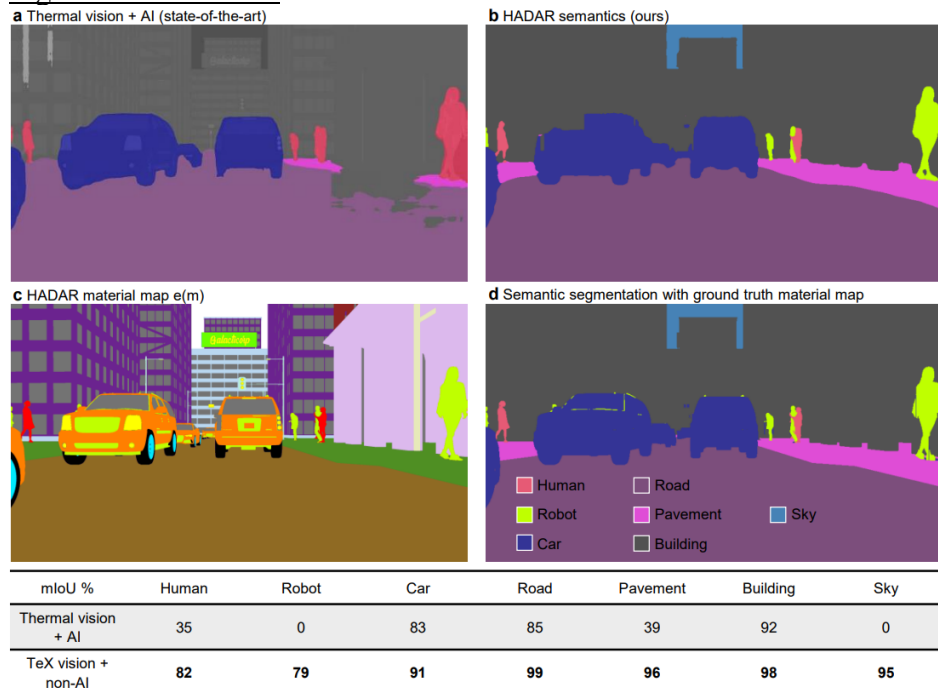


Fig.17 Comparison of our physics-driven semantic segmentation with the state-of-the-art AI-enhanced thermal semantic segmentation. HADAR semantic segmentation based on TeX vision beats the state-of-the-art ‘thermal vision + AI’ segmentation, showing HADAR efficacy in complicated scenes.

- Moreover, we want to clarify that our TeX vision or its material map (emissivity) itself is different from conventional semantic segmentation. In our approach, every material category has a discrete index label whereas in traditional semantic segmentation – every object has a discrete index label. We exploit the underlying physics that all materials have a spectral emissivity curve which arises from its causal

optical response function. HADAR requires a multi-spectral infrared thermal camera different from conventional panchromatic FLIR or optical cameras or LiDAR. TeX vision is an alternative physics-driven-representation of heat distinct from RGB vision or LIDAR point cloud. All other subsequent AI algorithms like optical/scene flow, stereo matching, semantic/instance segmentation, etc. that are previously developed on RGB vision can be adapted to TeX vision. We are excited to pursue these ideas through extensive future studies. Although, it is true that material map or emissivity is very similar to a semantic map. Material map is at the physical-component level, while semantic map functions at the object level. In this paper, we have used a heuristic non-machine-learning algorithm to transform the obtained material map into semantic segmentation to demonstrate HADAR efficacy. The algorithm is detailed in algorithm 4 of the Supple. Info. We tested our algorithms on our city block dataset, where we have multiple aluminum robots vs. humans, and aluminum is also used for car logo. Basically, each semantic category is a combination of several materials within the map, e.g., Car = window glass + tire + car paint + headlights + aluminum logo, Robot = aluminum, and so on. Neighboring pixel interactions have been used to transform the material map to semantic map, as shown in the above figure c and d. We emphasize that our heuristic semantic segmentation is not only based on emissivity but also neighboring pixels, and hence it won't fail if two subjects have the same emissivity. For example, car logo is aluminum, the same as robot, but they can be distinguished in our semantic segmentation (above Fig.b). In the future, our cityblock dataset can be used to design AI algorithms for semantic segmentation with TeX vision.

- At last, for people detection, we extract the material region corresponding to the desired target and only perform detection over the selected region, as shown below. For human detection, we extract the region of material 'human' from the material map, and the detection result gives bounding boxes of humans. For robot detection, we extract the region of material 'aluminum'. Even though car logo and other components of the car is also selected, people detection finds the correct spatial patterns and gives robot detection. Again, our approach won't fail even if two objects have the same materials. These results are added in Sec.SIIE of the Supple. Info.

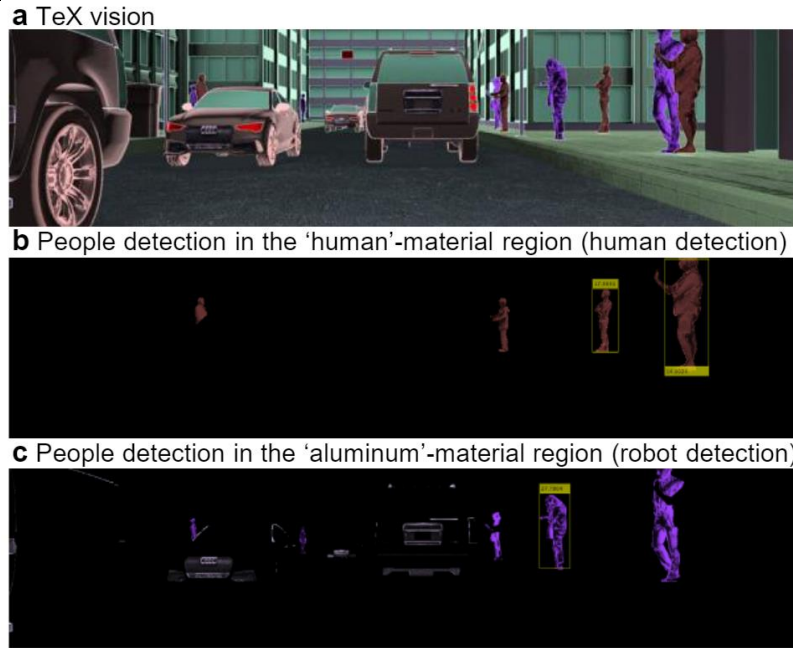


Fig.18 Demonstration of physics-driven object detection. HADAR can perform object detection over particular material regions, and hence HADAR can distinguish similar geometries with material signatures. This figure is given as Fig.S17 of the Supple. Info.

- We have also revised the caption of Fig.5 in the main text and relevant contexts from ‘physics-driven semantic segmentation’ to ‘physics-driven perception’ to avoid confusion.

Reference(s):

[1] Fu, Jun, et al. "Dual attention network for scene segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

C11 2. In the HADAR ranging, the authors claim that they have used several AI algorithms for instance, DeepPruner, PSMNet, but they have not provided any training details for these algorithms. Similarly, details of data collection and experimental results for these algorithms are not found in this manuscript. Qualitative and quantitative results would help readers and reviewers to make a fair comparison.

R11 We apologize that in the previous version of the manuscript DeepPruner and PSMNet are mentioned without explanation on the training data. In our work, we used pre-trained AI algorithms to demonstrate HADAR advantages over AI-enhanced thermal sensing. The reasons to use pre-trained models are three folds.
 Firstly, we note that in HADAR TeX vision, the scene is captured with physical attributes being represented by hue (material index), saturation (temperature) and value (texture). This novel representation has information content which is not present in the output of optical cameras (RGB vision), conventional IR thermal cameras (panchromatic thermal vision), or LiDAR (point cloud). Subsequent machine learning algorithms in computer vision regarding stereo matching, optical flow, scene flow, semantic segmentation, etc. that are previously based on RGB vision,

thermal vision or point cloud can be adapted to TeX vision. Developing new machine learning algorithms exploiting TeX vision presents a new research frontier and will be subject of future studies.

Secondly, our city block dataset and our outdoor experimental data are for demonstrating proof of concept TeX vision advantages over existing panchromatic thermal vision. We therefore use existing pre-trained NN models to compare both these representations of infrared thermal radiation.

Thirdly, it is the convention to test new data (TeX vision) with old models (pre-trained). This will reveal the features in new data compared to the old training data. For example, in the ‘Cats’ thermal dataset [1], they used pre-trained models to test stereo matching on their new data and show the challenge of thermal ranging, as cited below.

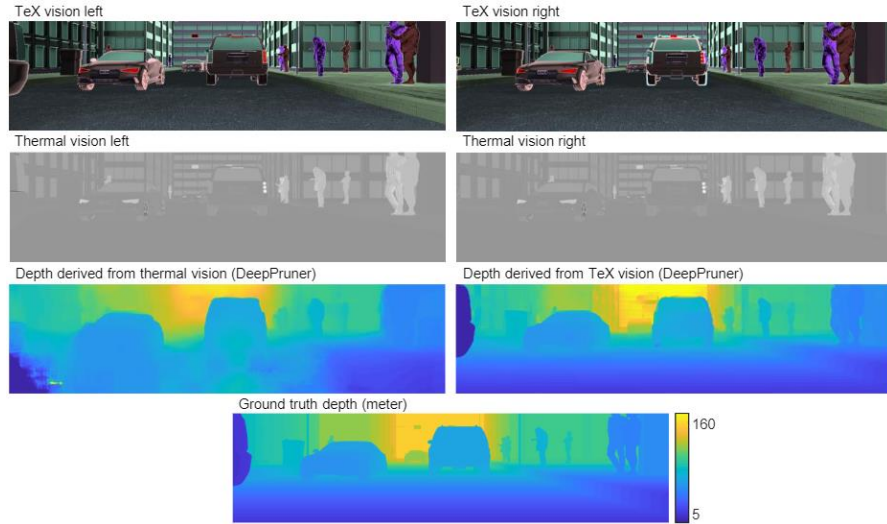
- MC-CNN [36] uses a convolutional neural network to learn a similarity measure between patches. We use two of the fast implementations, one trained on the Middlebury dataset [30], and one trained on the KITTI dataset [12].

To make fair comparisons between HADAR and AI-enhanced thermal sensing, we used DeepPruner (pre-trained on the KITTI dataset) on both thermal vision and TeX vision for stereo matching. Similarly, we used HOG+SVM (pre-trained on the INRIA Person dataset) on both thermal vision and TeX vision for people detection. Finally, we exploit DANet (pre-trained on the Cityscapes dataset) on thermal vision as a baseline comparison for the task of semantic segmentation. On the other hand, for HADAR semantic segmentation, we used non-machine-learning algorithms based on TeX vision. Above all, our comparisons between TeX vision and thermal vision are either characterized with the same AI algorithms, or with AI enhancement only for thermal vision, so that when HADAR outperforms AI-enhanced thermal sensing, the advantage is clearly from our TeX vision but not the algorithm.

To better show our comparisons, we have made the following revisions.

1. We have explicitly marked the training dataset when we introduce those AI algorithms.
 - In the caption of Extended Data Fig.5:
“...using HOG features and support vector machine pre-trained on the INRIA Person dataset...”
 - In the caption of Extended Data Fig.6:
“...using machine-learning-based DeepPruner (pre-trained on the KITTI dataset)...”
 - In the caption of Extended Data Fig.8:
“...thermal semantic segmentation with DANet (pre-trained on the Cityscapes dataset)...”
2. We have provided quantitative ranging comparisons in Extended Data Fig.6 and qualitative comparisons in Extended Data Fig.7. In comparing TeX vision with traditional thermal vision, we used DeepPruner as the state-of-the-art AI algorithm. The PSMNet we used in the previous version is pre-trained on the Scene Flow dataset. As its performance is not as good as DeepPruner and not representative, we have removed the PSMNet part in the new version. We have also replaced previous SGBM (semi-global block matching, non-machine-learning) results with the DeepPruner (machine learning) results in Extended Data Fig.7. Both non-

machine-learning and AI-based algorithms show TeX vision enhances the accuracy of ranging through texture recovery. New results of Extended Data Figs.6 and 7 are briefly cited below.



		Density	Mean error (px)	Accuracy (%)			
				$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 10$
Thermal vision + AI	Street	0.5	99.74	4.45	12.65	17.01	22.29
	Entire image	1	55.32	31.54	45.99	50.36	54.98
TeX vision + AI	Street	0.5	1.49	24.74	96.94	98.77	99.66
	Entire image	1	2.09	42.20	93.59	96.64	98.45

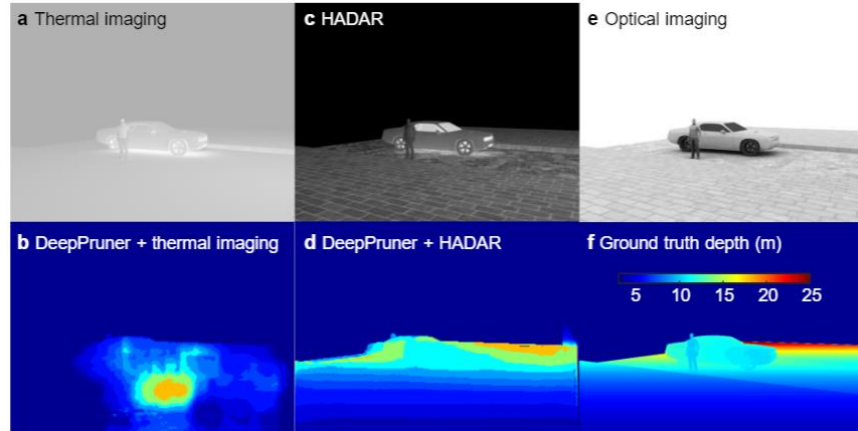


Fig.19 Quantitative and qualitative comparisons of ranging performances based on HADAR and traditional thermal vision. HADAR ranging beats state-of-the-art AI-enhanced thermal ranging. Top figure is the Extended Data Fig.6. Bottom figure is the Extended Data Fig.7

3. We have provided the details of our prototype HADAR calibration and outdoor data collection in ‘Methods --- prototype HADAR calibration and data collection’ and also in the current Extended Data Fig.9. We have also made our city block dataset public and available at https://drive.google.com/drive/folders/1da2Uh5t_QOy-MrWxhkJJw3MueNxsuVtn?usp=sharing; we will host it on Github for the scientific community once the paper is published) where we have described how the dataset is generated. The reason of using synthesized dataset and the procedures on how to generate an

	<p>experimental dataset have been detailed in Comment & Reply 3 (R3). The training details of our TeX-Net for TeX decomposition and TeX vision are given in Sec.SIIIA of the Supple. Info.</p> <p>Reference(s): [1] Treible, Wayne, et al. "Cats: A color and thermal stereo benchmark." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.</p>
C12	<p>3. "We develop HADAR estimation theory to address fundamental limits of object identification from its thermal infrared signature. We believe this will be crucial in guiding public policy for the industrial revolution where decision accuracy of machine perception can be bounded by physical laws as opposed to training data volume". It would help readers and reviews to appreciate the claims if an example or two are provided where the proposed technique helps public policy.</p>
R12	<p>We regret that the example we provided in the manuscript has not been explained clearly to make our claim easy to understand. The above claim about guiding public policies is related to the HADAR identifiability given by Eq.2 in the main text.</p> $I = \log_2 \left[1 + \operatorname{erf} \left[\sqrt{\frac{Nd_0^2}{2(1+\gamma)}} \right] \right], \quad (2)$ <p>HADAR identifiability, I ($0 \leq I \leq 1$), describes the maximum Shannon information of the target material that one can retrieve from N observed thermal photons. In the special case of two materials, $I=1$ signifies one bit information in the thermal photons causing the classification to always yield the correct material. For $I=0$, lack of information in the collected thermal signal causes random material classification i.e. 50% probability. Here, γ is a characterization of the used HADAR sensor (related to Noise-Equivalent Power or Special Detectivity), and d_0 is the semantic distance between a pair of candidate materials. The identifiable criterion (threshold) is given by $I_0 \approx 0.75$ ($\frac{Nd_0^2}{1+\gamma} = 1$), which means one can identify the target material if $I > I_0$, or $\frac{Nd_0^2}{1+\gamma} \geq 1$.</p> <p>The example we have shown is the human-robot identification problem in Fig.3 of the main text. The walking human-shaped target has two candidate materials, organic skin/fabrics or metallic aluminum, with semantic distance calculated to be $d_0 \approx 0.001$. This requires $\frac{N}{1+\gamma} \geq 10^6$ to identify the target if the environment is at $T_0 = 20 \text{ C}^\circ$ and $V_0 = 0.5$ (see Fig.3c). The observed photon number N is related to the human-robot scene, as well as the f-number (focal length f over the aperture size D), exposure time t, and pixel size A_p, see the heat signal model in Sec.I of the Supple. Info. Eventually, the above identifiable criterion leads to the minimum requirement of the hardware configurations, $\frac{tA_p}{(1+\gamma)(f/D)^2} \geq 5 \times 10^{-16}$. This minimum requirement of the hardware can be used to guide the public policies in the AI industry. For example, the lowest detectivity (or highest NEP), the smallest aperture size, the highest frame rate and hence the maximum travelling speed, etc., must meet the above inequality so as to be able to identify human vs. robot. If the detector doesn't meet the above requirement, its collected data will be insufficient in information. No matter how much data is collected and used to train</p>

	<p>a neural network (how large the training volume is), machine learning just cannot perform well (see the machine learning performance in Fig.3b of the main text when the photon number is insufficient, i.e., the normalized photon number is below 1).</p> <p>A second example is the other way around. If the detector is given, say, the FLIR A325sc camera, we have $\frac{tA_p}{(1+\gamma)(f/D)^2} = 8.16 \times 10^{-18}$ in one frame or shot. To have $I > I_0$, we must have $d_0 > 0.0078$, which means the FLIR A325sc camera can only distinguish sufficiently different material pairs in one shot, such as organic skin vs. glass mannequin ($d_0 = 0.049$). This minimum semantic distance identifiable by the given detector will also guide the public policies in the AI industry. For example, in which scenario the given camera can be used, and in which scenario the camera cannot. We note that the HADAR identifiability also applies to multi-material libraries, as can be seen in Sec.SIIB of the Supple. Info.</p> <p>To make our claim clearer, we have added the above details into the Method section ‘Guiding public policy’:</p> <p><i>“The HADAR identifiable criterion is $\frac{Nd_0^2}{1+\gamma} = 1$, which means one can identify the target material if $\frac{Nd_0^2}{1+\gamma} \geq 1$. The semantic distance between human body (skin) and robot (aluminum) in Fig.3 is calculated to be $d_0 \approx 0.001$. This requires $\frac{N}{1+\gamma} \geq 10^6$ to identify the target if the environment is at $T_0 = 20\text{ C}^\circ$ and $V_0 = 0.5$ (see Fig.3c). The observed photon number N is related to the human-robot scene, as well as the f-number (focal length f over the aperture size D), exposure time t, and pixel size A_p, see the heat signal model in Sec.SI of the Supple. Info. Eventually, the above identifiable criterion leads to the minimum requirement of the hardware configurations, $\frac{tA_p}{(1+\gamma)(f/D)^2} \geq 5 \times 10^{-16}$. This minimum requirement of the hardware will guide the public policies in the AI industry. For example, the lowest detectivity (or highest NEP), the smallest aperture size, the highest frame rate and hence the maximum travelling speed, etc., must meet the above inequality so as to be able to identify human vs. robot. If the detector doesn’t meet the above requirement, its collected data will be insufficient in information. No matter how much data is collected and used to train a neural network (how large the training volume is), machine learning just cannot perform well (see the machine learning performance in Fig.3b of the main text when the photon number is insufficient, i.e., the normalized photon number is below 1).</i></p> <p><i>If the detector is given, e.g., the FLIR A325sc camera, we have $\frac{tA_p}{(1+\gamma)(f/D)^2} = 8.16 \times 10^{-18}$ in one image frame. To meet the criterion, we must have $d_0 > 0.0078$, which means the FLIR A325sc camera can only distinguish sufficiently different material pairs in one image frame, such as organic skin vs. glass mannequin ($d_0 = 0.049$). This minimum semantic distance identifiable by the given detector will also guide the public policies in the AI industry. For example, in which scenario the given camera can be used, and in which scenario the camera cannot.”</i></p>
C13	<p>E. Suggested improvement</p> <p>In addition to the above comments, some additional comments are as follows that require more explanation.</p>

	1. “where multiple attributes are desired either for safety guarantees or scientific purpose”. Some examples with reference are required for this claim.
R13	<p>In this new version of our manuscript, we have cited the ‘phantom braking’ example to illustrate the argument ‘multiple physical attributes beyond visual appearance are desired for safety guarantees’ for autonomous navigation applications. With temperature and material identified, in addition to the visual shape, TeX vision can overcome the phantom braking issue by checking if the temperature or material of the phantom image is consistent with a real human body, as demonstrated with the Einstein cardboard in Fig.5 of the main text. This is also consistent with the spirit of safety redundancy in the multi-sensor fusion solution (cameras + sonar + radar + LiDAR) adopted in the self-driving industry to ensure safety (see, https://www.thedrive.com/tech/17541/heres-how-nvidia-plans-to-ensure-self-driving-car-safety, and also [1]).</p> <p>As for wildlife monitoring, we have added a reference [2] to help readers appreciate the application. In the reference of ‘multimodal wireless sensor network for wild life monitoring’, optical and thermal cameras are used to obtain different attributes (shape + temperature) and perform monitoring during the whole day as well as night time. Shape with geometric textures is better suited for recognition, while temperature pattern is better suited for analyzing health conditions of wildlife. However, most wild animals are active at night when optical cameras don’t work. In this scenario, TeX vision recovering temperature as well as geometric textures can provide both valuable attributes of shape + temperature close to the fundamental bound of precision.</p> <p>Accordingly, the relevant part is revised as below: “Major advantages of TeX semantics will be found in autonomous navigation and wildlife monitoring, where multiple physical attributes beyond visual appearance are desired either for safety guarantees <u>[#1]</u> or scientific purposes <u>[#2]</u>.”</p> <p>Reference(s): [1] Khaleghi, Bahador, et al. "Multisensor data fusion: A review of the state-of-the-art." Information fusion 14.1 (2013): 28-44. [2] Lopes, Carlos Eduardo Rodrigues, and Linnyer Beatrys Ruiz. "On the development of a multi-tier, multimodal wireless sensor network for wild life monitoring." 2008 1st IFIP Wireless Days. IEEE, 2008.</p>
C14	2. “However, large scale temperature screening with existing noncontact infrared thermometer or infrared thermography is ineffective due to lack of adaptivity to emissivity (complexion /makeup), distance, age, gender, and circadian variations [36–38].” A brief comment on the utility of TeX, in this case, will help the reader to appreciate.
R14	We thank the reviewer for the question as we believe TeX vision will have an important role to play for reaching the fundamental bound of temperature estimation through a completely non-contact approach. Traditionally, the thermal infrared cameras do not capture spectral information and therefore the estimate of temperature or emissivity is biased and far from the fundamental precision bound. In HADAR, we exploit spectral resolution to analyze and distinguish different emissivities (complexion/skin variabilities) with different spectral features. One major roadblock to achieving the fundamental bound of temperature estimation is that traditional panchromatic cameras and algorithms do not account for X i.e. the environmental thermal emission which

	<p>enters the camera and completely changes the temperature estimate. In TeX vision, the spectral resolution along with the HADAR constitutive equation helps us separate X (environmental factors) and reach the fundamental bound of temperature estimation in the image.</p> <p>In follow-up research, we are proposing to also build the material library for different human complexions and human skin variabilities. Consequently, HADAR with TeX vision can identify emissivity and recover textures for age, gender, and circadian recognitions. In addition to HADAR ranging based on TeX visions, HADAR is promising in adaptivity to emissivity (complexion/skin variabilities), distance of the target, age, gender, and circadian variations. In this new version of manuscript, we have added a brief comment to further explain how HADAR and TeX vision could help temperature screening: “However, large scale temperature screening with existing noncontact infrared thermometer or infrared thermography is ineffective due to lack of adaptivity to emissivity (complexion/skin variabilities), age, gender, circadian variations and distance of the target [37–39]. <u>As illustrated above, HADAR with TeX vision can identify spectral emissivity, estimate distance, and recover textures, promising in advanced adaptivity for more accurate temperature estimation.</u> Here, we experimentally demonstrate that HADAR thermography can automatically recognize emissivity (also with ranging) and reach the Cramer-Rao bound on temperature accuracy”</p>
C15	<p>3. “Cramér-Rao bound is therefore promising for the smart healthcare industry including early reliable skin cancer detection.” Reference and a brief comment will help users to understand the relation.</p>
R15	<p>Early skin cancer detection via thermography is a cutting-edge research direction [1-3]. The key idea is that tumor cells are more active and are of higher temperature than regular cells. The temperature difference could be as high as 0.25 Celsius degree, as reported in [4]. However, the signal captured by a thermal camera is the radiance S that includes scattering contributions from the environment (X) along with direct emission from the cancerous cells. Having a hot object (other people, instruments) in the patient room (or, considering X or not) makes a striking difference in estimated temperatures. We start with the HADAR constitutive equation</p> $S_{\alpha\nu} = e_{\alpha\nu}B_{\nu}(T_{\alpha}) + [1 - e_{\alpha\nu}]X_{\alpha\nu}, \quad (1)$ <p>As an example, the emissivity of skin can be well approximated as a constant of 0.95, and we assume that the environment is a blackbody ($X=B$) to approximately see the errors arising from ignoring the environment. The presence/absence of environmental scattering signal (X) is equivalent to a 5% relative difference of direct emission of the target ($B(T_{\alpha})$), which corresponds to 3 Celsius degree temperature variation around the standard 37 Celsius degree temperature. This error arising from ignoring the environmental signal is much larger than the temperature difference caused by tumor cells. To minimize this effect, K. Tang et. al. in [4] performed experiments ‘either in an open-area, outdoor environment under clear sky (cloud free), or using a cold-plate setup’, which restricts the indoor applications for fever surveillance. Since TeX vision decomposes S, HADAR can reach the Cramer-Rao bound of temperature in the entire scene by accurately estimating e and X. This approach is promising for reliable skin cancer detection from thermal infrared images which have spectral resolution and TeX vision. Accordingly, we have added a reference in the main text:</p>

	<p>“Cramér-Rao bound is therefore promising for the smart healthcare industry including early reliable skin cancer detection [1]” and we have added the above details to the Method section ‘HADAR thermography’ to make our claim clearer: “... <i>The temperature difference between tumor cells and regular cells in skin cancer could be as high as 0.25 degree Celsius. However, the signal captured by a thermal camera is the radiance S that includes scattering contributions from the environment (X) along with direct emission from the tumor cells. Having a hot object (other people, instruments) in the patient room (or, considering X or not) striking difference in estimated temperatures. As an example, the emissivity of skin can be well approximated as a constant of 0.95, and we assume that the environment is a blackbody ($X=B$) to approximately see the errors arising from ignoring the environment. The presence/absence of X is equivalent to a 5% relative difference of $B(T)$, which corresponds to 3 Celsius degree temperature variation around the standard 37 Celsius degree temperature. This error arising from ignoring the environmental signal is much larger than the temperature difference caused by tumor cells. To minimize this effect, accurate thermography is limited to ‘either an open-area, outdoor environment under clear sky (cloud free), or using a cold-plate setup’, which restricts the indoor applications for fever surveillance. Since TeX vision decomposes S, HADAR can reach the Cramer-Rao bound of temperature by properly estimating e and X and hence is promising for reliable skin cancer detection”</i></p> <p>Reference(s): [1] Magalhaes, Carolina, et al. "Comparison of machine learning strategies for infrared thermography of skin cancer." <i>Biomedical Signal Processing and Control</i> 69 (2021): 102872. [2] Magalhaes, C., Ricardo Vardasca, and J. Mendes. "Recent use of medical infrared thermography in skin neoplasms." <i>Skin Research and Technology</i> 24.4 (2018): 587-591. [3] Iljaž, J., et al. "Solving inverse bioheat problems of skin tumour identification by dynamic thermography." <i>Inverse Problems</i> 36.3 (2020): 035002. [4] Tang, Kechao, et al. "Millikelvin-resolved ambient thermography." <i>Science advances</i> 6.50 (2020): eabd8688.</p>
C16	<p>4. “Our results call for heat exploitation in the quantum regime where single photon detectors are being developed in the thermal infrared”. It is not mentioned in the whole script except in the introduction, a bit of explicit comment may help the readers.</p>
R16	<p>We agree. We would like to clarify that conventional cameras (visible or infrared) are not sensitive enough to measure single particles of light i.e. single photons. In the visible spectral range, there exists single photon avalanche detector arrays and EMCCDs that have single photon sensitivity which is the fundamental limit for a detector’s sensitivity. However, such detectors do not exist for the thermal mid-infrared and long-wave infrared spectral range. Only recently has research work in superconducting detectors [1] begun to address this urgent need. We believe such new class of single photon detectors in the infrared spectral range can lead to new frontier of applications of HADAR + TeX vision. This is true since our theory works in the shot noise limited regime which is the boundary of the performance regime between classical IR thermal cameras and quantum single photon detector arrays. Our results of shot-noise limits to detection and ranging can be compared to these new emerging quantum detectors once they are used in heat exploitation.</p>

	<p>Accordingly, we have revised the above statement to help multi-disciplinary readers understand the connection: “<u>Our shot-noise limits of detection and ranging set the benchmark and</u> call for heat exploitation in the quantum regime where single photon detectors are being developed beyond visible spectral range into the thermal infrared [1]”</p> <p>Reference(s): [1] Reddy, Dileep V., et al. "Superconducting nanowire single-photon detectors with 98% system detection efficiency at 1550 nm." Optica 7.12 (2020): 1649-1653.</p>
C17	<p>5. “However, large scale temperature screening with existing non-contact infrared thermometer or infrared thermography is ineffective due to lack of adaptivity to emissivity (complexion/makeup), distance, age, gender, and circadian variations' '. Reference and a brief comment will help users to understand the relation.</p>
R17	<p>We have added a comment as explained in Reply 14 (R14). “However, large scale temperature screening with existing noncontact infrared thermometer or infrared thermography is ineffective due to lack of adaptivity to spectral emissivity (complexion/skin variabilities), distance of target, age, gender, and circadian variations [37–39]. <u>As illustrated above, HADAR with TeX vision can identify spectral emissivity, estimate distance, and recover textures, promising in advanced adaptivity for more accurate temperature estimation.</u> Here, we experimentally demonstrate that HADAR thermography can automatically recognize emissivity (also with ranging) and reach the Cramer-Rao bound on temperature accuracy”</p>
C18	<p>6. They mention a “phantom breaking phenomenon” as a disadvantage of thermal imaging, but do not explain if the proposed technique addresses it.</p>
R18	<p>We regret that we haven’t clearly explained in the previous version that HADAR can address the phantom braking problem. As demonstrated in Fig.5a-c of the main text, the Einstein cardboard is misunderstood as a human body by optical cameras (a) and LiDAR (c), however, it is clearly distinguished from a human body by HADAR (b). HADAR TeX vision can check if the temperature and material of a phantom image are consistent with a real human body. This is also demonstrated in the summer experiment in Extended Data Fig.10. Extended Data Figs.5 and 8 where robots are distinguished from the human bodies also demonstrate our argument. Accordingly, we have revised it in the section of HADAR semantics as below: “<u>We now experimentally demonstrate HADAR in an outdoor scene and illustrate how it addresses phantom braking ... HADAR detects people only in the corresponding material region (skin+fabrics) and clearly distinguishes it from the cardboard, providing an approach to overcome the phantom braking problem [ref].</u>”</p>
C19	<p>F. References 1. “The emerging Industry 4.0 of smart technologies [18] calls for a future with scalable human-robot social interactions since it is expected that one in ten vehicles will be automated by 2030 and 100 million robot helpers will be serving people.” The reference paper has no such claim.</p>

R19	<p>We apologize that references are not properly given in the previous version and it caused confusion. The previous Ref. [18] is solely given as a review of smart technologies. The statistical projection of ‘one in ten vehicles will be automated by 2030’ can be found at https://mailchi.mp/statista/autonomous_cars_20200206?e=145345a469. There are two versions of statistical projections of the number of robot helpers by 2030. One clearly says 20 million robots (https://resources.oxfordeconomics.com/how-robots-change-the-world) while the other says 800 million jobs will be taken by robots (https://www.bbc.com/news/world-us-canada-42170100). We do anticipate that statistical projections might not be accurate. Previously, we took a number in-between, but now we realize it is better to stick to the relatively robust and original data. Therefore, we have revised 100 million to 20 million and added these two links: “The emerging Industry 4.0 of smart technologies [18] calls for a future with scalable human-robot social interactions since it is expected that one in ten vehicles will be automated by 2030 [16] and <u>20 million robot helpers</u> will be serving people [17].”</p>
C20	<p>2. “Scalable perception”. No explanation is given for the scalable perception</p>
R20	<p>We use ‘Scalable perception’ to denote the machine perception techniques that can support simultaneous operations of multiple intelligent agents (IA). For example, active modalities like sonar, radar, and LiDAR can work accurately on one single IA, but they will immediately have signal interference issues when two or more IA’s work together simultaneously. ‘Eye safety’ is the main restriction to the detection range of LiDAR (tens or hundreds of meters). However, if there are N=100 self-driving cars using LiDAR on the same street, the signal emission power should be decreased further by N=100 folds, to ensure eye safety. Technically, large number of agents (N) will decrease the ranging distance of every self-driving car due to the lower power emission budget. Therefore, we mentioned that active modalities face the key challenge of scalability.</p> <p>Accordingly, we have revised line 6, second column, page 1 to define the concept of scalable perception:</p> <p>“However, <u>simultaneous perception of the scene by numerous agents (scalable perception) is fundamentally prohibitive</u>”</p> <p>We have also added the scaling law and eye safety in a footnote:</p> <p>“[11] J. Hecht, Lidar for self-driving cars, Opt. Photon. News 29, 26 (2018). <u>Eye safety requires the emitting power of an agent to scale down as the inverse of the number of agents</u>”</p>
C21	<p>3. They claimed that this method is novel but the following are the works that have done temperature emissivity separation.</p> <p>(1) Jie Cheng, Qing Xiao, Xiaowen Li, Qinhuo Liu, Yongming Du, Aixiu Nie, "Multi-layer perceptron neural network based algorithm for simultaneous retrieving temperature and emissivity from hyperspectral FTIR dataset", Geoscience and Remote Sensing Symposium 2007. IGARSS 2007. IEEE International, pp. 4383-4385, 2007.</p> <p>(2) Xinghong Wang, Xiaoying OuYang, Bohui Tang, Zhao-Liang Li, Renhua Zhang, "A New Method for Temperature/Emissivity Separation from Hyperspectral Thermal Infrared Data", Geoscience and Remote Sensing Symposium 2008. IGARSS 2008. IEEE International, vol. 3, pp. III - 286-III - 289, 2008.</p>

	<p>(3) Hang Yang, Lifu Zhang, Junyong Fang, Xia Zhang, Qingxi Tong, "Algorithm research of building materials emissivity extracting", Geoscience and Remote Sensing Symposium (IGARSS) 2010 IEEE International, pp. 3350-3353, 2010.</p> <p>(4) Hang Yang, Lifu Zhang, Li Liu, Qingxi Tong, "Temperature and emissivity separation from TASI data based on wavebands selection", Geoscience and Remote Sensing Symposium (IGARSS) 2011 IEEE International, pp. 1850- 1853, 2011.</p> <p>(5) Ning Wang, Yonggang Qian, Hua Wu, Lingling Ma, Zhao-Liang Li, Lingli Tang, "Performances of temperature and emissivity separation methods for hyperspectral thermal data affected by the changes of spectral properties of sensor", Geoscience and Remote Sensing Symposium (IGARSS) 2013 IEEE International, pp. 2152-2155, 2013.</p> <p>(6) Schmugge, Thomas, Andrew French, Jerry C. Ritchie, Albert Rango, and Henk Pelgrum. "Temperature and emissivity separation from multispectral thermal infrared observations." Remote Sensing of Environment 79, no. 2-3 (2002): 189-198.</p> <p>(7) V. Payan Corresponding author & A. Royer (2004) Analysis of Temperature Emissivity Separation (TES) algorithm applicability and sensitivity, International Journal of Remote Sensing, 25:1, 15-37, DOI: 10.1080/0143116031000115274</p>
R21	<p>We regret that the novelty of our TeX decomposition has not been explained clearly in the previous version. Here, we would like to clarify our novelty and make comparisons with the above references.</p> <p>One of the motivations underlying our work of HADAR is the ‘ghosting effect’ challenge in existing thermal imaging which cannot be tackled by the conventional approaches to temperature-emissivity (TE) separation. To recover the (geometric) texture and overcome the ‘ghosting effect’, we analyzed the full heat signal model as shown below and proposed the TeX decomposition which immediately leads to TeX vision.</p> $S_{\alpha\nu} = e_{\alpha\nu}B_{\nu}(T_{\alpha}) + [1 - e_{\alpha\nu}]X_{\alpha\nu}, \quad (1)$ $X_{\alpha\nu} = \sum_{\beta \neq \alpha} V_{\alpha\beta}S_{\beta\nu}.$ <ol style="list-style-type: none"> 1. The key to our HADAR theory is the consideration of texture X with the correct spectral and scene-dependent structure. In the mathematic structure of X, unknown thermal lighting factors V depict local surface normal of the target and hence is crucial to overcome the ‘ghosting effect’ and improve ranging. This leads to rich texture information in X, instead of a spatially uniform constant, and enables TeX vision. 2. In contrast, X is either ignored in the literature or approximated as a constant spectrum without a structure, incapable of revealing the local geometric texture (surface normal). As we have cited in the main text before, the TE separation is firstly proposed in Ref.[1]. Gillespie et. al. in [1] approximated X as a given downwelling atmospheric irradiance S_{\downarrow} as in their heat signal model L $L \approx \tau\varepsilon B(T) + \tau\rho S_{\downarrow} + S_{\uparrow}.$ <p>where $\rho = 1 - e_{\alpha\nu}$, their ε is our $e_{\alpha\nu}$, and τ and S_{\uparrow} are irrelevant to this comment. Since $X_{\alpha\nu} = S_{\downarrow}$ is assumed known and independent of target α, it becomes a uniform term ready to subtract and cannot capture the local surface normal of the target. Therefore, the inverse problem is simplified from TeX decomposition to TE separation without X. We emphasize</p>

that the explicit structure of X in our HADAR theory complicates the inverse problem quite a lot but brings us the benefit of texture.

3. All other TE-separation papers mentioned by the reviewer follow the same heat signal model used in [1] that cannot capture local texture. Explicitly,

- The above reference (1) pointed out that the TE separation proposed in [1] doesn't work well under certain conditions when the difference of ground-leaving radiance and object's blackbody radiation at its true temperature and the instrument random noise are on the same order. They proposed a 3-layer perceptron neural network to settle the defect. However, reference (1) uses spatially uniform S_{\downarrow} and cannot capture local texture.

- The above reference (2) follows a similar heat signal model as [1], see their equation 2,

$$L_i^{surf} = \varepsilon_i B_i(T) + (1 - \varepsilon_i) L_i^{\downarrow} \quad (2)$$

Reference (2) tried to improve the estimation of T and e by making use of spectral features of downwelling atmospheric irradiance L_{\downarrow} . But again, reference (2) uses spatially uniform L_i^{\downarrow} and cannot capture local texture.

- The above reference (3) analyzed and compared different algorithms (iterative vs. non-iterative) for TE separation. It follows a similar heat signal model as [1], see their equation 2,

$$L_j = \varepsilon_j \tau_j B_j(T_s) + \tau_j (1 - \varepsilon_j) L_{atm,j}^{\downarrow} \quad (2)$$

But again, reference (3) uses spatially uniform $L_{atm,j}^{\downarrow}$ and cannot capture local texture.

- The above reference (4) analyzed the SNR of each band and the temperature accuracy based on wavebands selection. Reference (4) follows a similar heat signal model as [1], see their equation 6,

$$L_j = \varepsilon_j \tau_j B_j(T_s) + \tau_j (1 - \varepsilon_j) L_{um,j}^{\downarrow} \quad (6)$$

Again, reference (4) uses spatially uniform $L_{atm,j}^{\downarrow}$ and cannot capture local texture.

- The above reference (5) analyzed the robustness and accuracy of several algorithms of TE separation, under different instrument characteristics, including the shifting of spectral and the broadening of the full-width half-maximum (FWHM). It follows a similar heat signal model as [1], see their equation 1b,

$$L_{ag}(\lambda) = \varepsilon(\lambda) B(\lambda, T) + (1 - \varepsilon(\lambda)) R_{at\downarrow}(\lambda), \quad (1b)$$

Again, reference (5) uses spatially uniform $R_{at\downarrow}$ and cannot capture local texture.

- The above reference (6) presented data from their TIMS (Thermal Infrared Multispectral Scanner) instrument and applied traditional TE separation algorithm to process the data. Their heat signal is like [1], see their equation 3,

$$L_j(surf) = \varepsilon_j L_{BB}(\lambda_j, T_{grd}) + (1 - \varepsilon_j) L_j(atm \downarrow) \quad (3)$$

Again, reference (6) uses spatially uniform $L_j(atm \downarrow)$ and cannot capture local texture.

- The above reference (7) is to assess the TE separation proposed by [1], showing the applicability of TE separation for hyperspectral data. The heat signal model is consistent with [1], see their equation 5

$$L_{\text{atm},j}(\theta_r, \varphi_r) = \varepsilon_j(\theta_r, \varphi_r) B_j(T_s) + (1 - \varepsilon_j(\theta_r, \varphi_r)) L_{\text{atm},j} + \rho_{\text{e},j}(\theta_s, \varphi_s, \theta_r, \varphi_r) E_{\text{atm},j}(\theta_s) \quad (5)$$

Again, reference (7) uses spatially uniform $L_{\text{atm},j}$ and cannot capture local texture.

- In TeX decomposition, we use a material library for dimensional reduction. This is naturally suitable for material classification or semantic segmentation in intelligent applications. In contrast, TE separation usually assumes lower and upper bounds of spectral emissivity to estimate temperature and subsequently estimate emissivity. The estimated emissivity in TE separation itself doesn't have implication in semantic segmentation i.e. there is no mapping to the specific physical material within a library.
- The consideration of texture X is also crucial for accurate estimation of temperature T and material e . In the literature of thermal imaging, the texture X term is usually ignored, see e.g., [1] (K. Tang et al use the Stefan-Boltzmann law which is an integral of the self-emission term $e \cdot B$, without considering reflection from X). As mentioned in Comment & Reply 15, the ignorance of X can lead to up to 3 Celsius degree temperature uncertainty. Furthermore, in our outdoor experiment, we have also confirmed that correct material classification (estimation of e) is only possible with the TeX model rather than the traditional temperature-emissivity model (without X term), as shown below.

Material classification with/without TeX

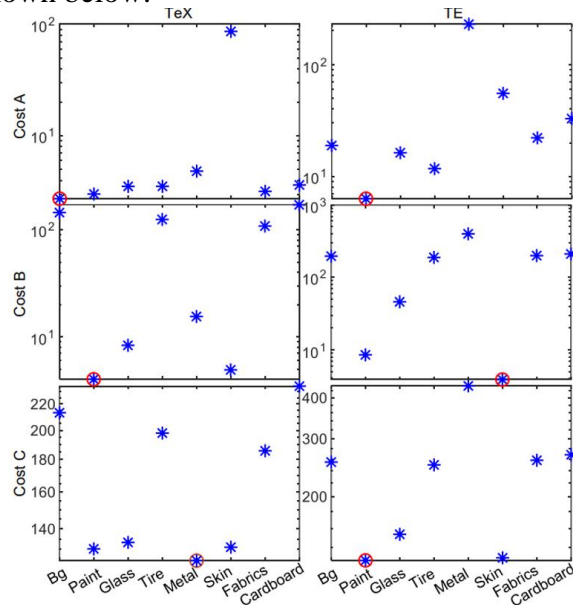
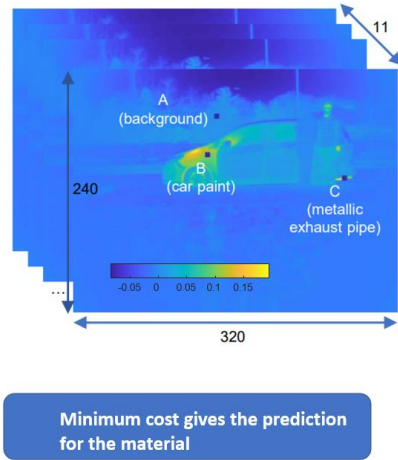


Fig.20 X term is also crucial for material classification (e estimation). For three sample pixels, A, B, and C, in the left image, TeX model gives the correct material classification, while TE model returns wrong classification. Cost (blue stars) is defined as the residual error of least-squares fitting to TeX or TE models. The minimum cost (red circles) gives the prediction for the material.

Above all, the TE separation in the literature assumes spatially uniform X (no mathematic structure of X that correctly captures local surface normal) or simply ignores X , leading to

	<p>the loss of textures and inaccurate temperature and material estimation. Technically, TE separation is only a simplified model of our TeX decomposition. We believe our TeX decomposition is a novel perspective in understanding the heat signal, especially in recovering textures and overcoming the ‘ghosting effect’.</p> <p>Accordingly, we have revised Methods Section ‘TeX decomposition’ (last line) to explain and emphasize this argument:</p> <p><i>“...We verify that TeX decomposition is crucial for vision applications and goes beyond the traditional TE (temperature-emissivity) separation approach, see Extended Data Fig.13. We emphasize that TE separation completely ignores the environmental heat processes or assumes spatially uniform environmental heat signal (X). In stark contrast, TeX decomposition captures the interplay between the complex real-world scene and its non-uniform environment through the HADAR constitutive equation (Eq 1). Thus TeX decomposition captures local surface normals of objects in the scene that arises from environmental thermal illumination and carries crucial information about texture.”</i></p> <p>Also, we have added the above material classification result with/without the X term in <u>Extended Data Fig.13</u> to show the importance of X term in analyzing heat signal.</p> <p>Furthermore, <u>we have also added the influence of X on temperature estimation to the last paragraph of Method section ‘HADAR thermography’.</u></p> <p>Reference(s): [1] Tang, Kechao, et al. "Millikelvin-resolved ambient thermography." Science advances 6.50 (2020): eabd8688. [2] A. Gillespie, S. Rokugawa, T. Matsunaga, J. S. Cothorn, S. Hook, and A. B. Kahle, A temperature and emissivity separation algorithm for advanced spaceborne thermal emission and reflection radiometer (aster) images, IEEE Trans. Geosci. Remote Sens. 36, 1113 (1998).</p>
C22	<p>G. Clarity</p> <p>1. Language use in writing is a bit extreme (e.g. “that can disrupt AI industry”, “TeX degeneracy”)</p>
R22	<p>We thank the reviewer for the suggestions. We have removed extreme words sticking to objective statements.</p> <p>But, we still kept TeX degeneracy, as in the physics community, the word ‘degeneracy’ is used when multiple states of the system (T,e,X configurations) correspond to the same observable (S). Instead, we have explicitly pointed this out in Methods --- TeX degeneracy:</p> <p><u>“...Physical states having distinct triplet of TeX attributes but having the same observed heat S is addressed as TeX degeneracy.”</u></p>
C23	<p>2. The authors refer to the supplementary information frequently but do not mention the section which becomes bothersome for the reader</p>
R23	<p>We apologize that the previous version is not well organized and properly referenced. In this new version, we have expanded the Methods and Supple. Info., and clearly cited the Section number or Figure number whenever mentioned elsewhere.</p>
C24	<p>H. Decision</p> <p>1. The proposed technique is a fascinating idea and can benefit the AI researchers with another reliable sensor.</p>

	2. However, the current article is not ready for publication in its current form. It is suggested to consider the recommendations and resubmit.
R24	We are glad that the reviewer appreciates our proposed technique. We thank the reviewer again for the time and efforts spent to provide us with such detailed comments. We have addressed each comment individually above and made major revisions to significantly improve the quality of the manuscript. With these changes, we believe the current version is now ready for publication.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The reviewer recognizes and appreciates the diligence and effort shown by the authors in addressing the provided comments and concerns. While the authors have done a great job in laying out the theoretical framework and have provided sufficient mathematical and physics-based explanations for their approach, some major issues remain concerning the scientific soundness of the proposed ML/DL approaches and their correctness. The work certainly has the potential to present a breakthrough in the field of computer vision and AI. However, the majority of the testing is done on synthesized images but makes claims that it can outperform models that are largely applied to real-world imagery and the noise that such imagery inevitably encompasses. To justify these claims, the reviewer's greatest concern is that the synthetic datasets used by the authors are not comparable to real-world datasets acquired by HADAR. Despite the novel hardware design, the authors pursue to majorly utilize the synthetically generated datasets which cripple the proposed novel design. To truly prove the claim of superior performance and provide the breakthrough alluded to in the paper, the reviewer needs to see similar performance when using real-world datasets as opposed to results-based majorly on synthetic datasets.

The following items present the remaining concerns of the reviewer which to be addressed before the article can be accepted:

- 1) The reviewer acknowledges the work done to create the new spectro-spatial deep learning model presented in the paper compared against the multi-layer perceptron as presented in the earlier version of the manuscript. Still, there are some concerns about the generalized performance of the proposed model for multiple scenarios where the scenes depicted, vary as is typical of real-world imagery. So far in the paper, only select scenes have been presented where the HADAR model beats the SOTA models. It is equally important to gauge the model's performance over multiple images that are dissimilar and report the overall performance on these datasets.
- 2) The reviewer is keen to have a look at the TexNet code to examine the parameters and their efficacy. It is not clear if the performance reported in the paper is for one specific test set. The inclusion of cross-fold validation results on the dataset would eliminate the bias in test performance.
- 3) It is not clear how the authors generate Texvision from T, E, X components. Is the TexNet framework trained end to end? Specifically, since the TexNet is trained with a physics-based loss on a different objective, how are the authors able to realize Texvision out of the learned model? It is very important to elaborate these details to the readers as this is one of the fundamental strengths of this paper.
- 4) The authors are presenting their HADAR framework as a real-world breakthrough as opposed to a theoretical design and attempt to demonstrate real-world efficacy. However, the results presented are largely based on synthetic images. For example, Figures 1, 2,3,5,6,7,8, &11 are all based on synthetic images and correspond to results tables. In contrast, Figures 4, 10, & 12 are from a real-world image but it is the same image and none of these figures correspond to results tables showing

model efficacy. Thus, the reporting on comprehensive model performance is centered on images that do not accurately represent the noise found in real-world data and thus can not be representative of real-world efficacy, only validation of theoretical concepts. In other words, there exists a huge domain gap between the presented synthetic LWIR images and the HADAR captured images. While the synthetic images are valuable for some demonstration, the reviewer suspects model performance was gauged solely on synthetic images and thus the results presented do not reflect real-world performance. This is a major issue with the paper.

5) Although the authors tested the figures with the RGB image pre-trained people detection framework, the domain gap between source and target image domains would lead to incorrect detection for target images. Be careful that although the framework correctly identifies humans for this specific frame, a generalized performance estimation over multiple frames would be useful. Still, it is recommended to train the detection model with target images with few-shot learning to avoid the requirement for large labeled datasets. Also, be informed that the thermal and the HADAR results don't have significant texture information to correctly estimate the HOG features to be able to perform detection on these features. This is also evident with the detection score as shown in extended data fig5. Alternately, you could use SOTA human detection frameworks designed for thermal images. Take into consideration that the detection framework suffers if the HOG features generated for both the Robot and humans are the same since the only difference between these is the semantic color and some textures.

6) As the authors pointed out, the identification of the object is dependent on the semantic/statistical distance of each material in the library, which decreases when materials are introduced to the library. To address this issue, the authors propose to use a setup with higher multispectral band capabilities but their hardware itself is constrained in that it cannot identify even a minimal number of bands that occur in a real-world application. Firstly, these sorts of hardware constraints are not considered in synthetic examples which are used to demonstrate the superior performance on various tasks compared to thermal or other modalities. Secondly, realizing these constraints, I am skeptical of their performance in real-world situations where design costs are the bottleneck. The authors emphasize the need for a low-cost design for HADAR optimal design, but this will not allow the design to function optimally for optimal setup for separability of materials in the real-world situation. In the synthetic examples to prove HADAR efficacy, these constraints are relaxed as a result of which the authors can demonstrate many sets of materials whereas, for the real world, the material library is a bare minimum and insufficient to capture the variation in spectral signatures encountered in the real world.

Please carefully address these concerns for the next submission.

Referee #3 (Remarks to the Author):

In this paper, the authors proposed and demonstrated HADAR(Heat-Assisted Detection and Ranging) for fully-passive and physically-aware machine perception. This work is interesting, the authors exploited physics-driven perception to achieve improved performance against AI-enhanced thermal sensing. The paper is not easy to read, and I suggest authors should move some essential information from supplementary material to the article. Following are a few suggestions and some

questions for the authors:

1. Although more complex data are used in experiments suggested by other reviewer, I still feel the scenes visualized in the article are not complex enough. I hope the authors will add more complex scenes such as dense crowds and dense vehicles in the additional content.

2. The authors mention that HADAR is expected to make great progress in industry, but it seems that there is no demonstration of HADAR computational efficiency and deploy-ability in the article.

3. In HADAR, a deep neural network approach is used to predict materials by spectra, and I'm rather curious about what the results will be for HADAR if the prediction is wrong in this process. Is the prediction of materials by spectra robust enough?

4. In the comparison of this paper, the authors' distinction between humans and robots does not look extensive enough and there should be more visual examples to support the effectiveness of the proposed approach.

5. The experiments need improvement. In the paper, the algorithm of Thermal sensing + AI comparison is not state-of-the-art, and the authors should add the comparison with the latest algorithm.

6. Some experimental results seem unconvincing. For example, in Extended Data Fig. 8, why choose 10 frames of the left camera in the city block dataset 1? Can you evaluate them on entire dataset?

7. The fusion of infrared image and visible image perception is a common practice in real scenes. The authors should compare the performance after fusing optical imaging with raw thermal vision and HADAR TeX vision.

8. In open environments, there exist unknown classes of materials and unknown scenes. I wonder how the TeX vision would work in open environments.

Author Rebuttals to First Revision:

Cover letter to Reviewer 1

We would like to thank the reviewer for the encouraging response and very insightful and valuable comments. It certainly helped us to significantly revise our manuscript. In this round of review, there is one remaining major concern from all reviewers. Here, we briefly list all the major revisions, and we will provide individual replies to each comment from the next page onwards.

Major concern: Will HADAR work in a real-world environment?
The reviewers want to see (i) real-world HADAR performance and (ii) generalized performance over multiple dissimilar scenes. Explicitly, the reviewers want to see the performance in more complex synthetic scenes such as dense crowds and dense vehicles, and in real-world open environments with inevitable sensor noise and unknown classes of materials.
Major revision(s): Real-world demonstration of HADAR
In this latest revised manuscript, we have made major revisions to fully address the concerns.
<ol style="list-style-type: none"> We have added HADAR prototype-2 and real-world experiments, in the presence of sensor noise and unknown materials. Based on the input from the reviewer and to respond to the detailed questions, we formed a partnership with DARPA (The Defense Advanced Research Projects Agency, through the Invisible Headlights project) and the Army night-vision team (Infrared Camera Technology Branch, DEVCOM C5ISR Center, U.S. Army). We have now collected real-world experimental data using a pushbroom hyperspectral imager (~\$1M) and it took ~\$20K a day for personnel to collect data. Along with our previous home-built HADAR prototype-1 device, we call this as HADAR prototype-2. This data will be made available to the global research community accelerating progress not only in machine learning algorithms but also in the creation of new cost-effective and cheap sensors for HADAR. We have also generalized our HADAR theory so that it does not require an input of material library. We have created a HADAR database with 11 dissimilar scenes to test generalized HADAR performance. Our HADAR database consists of complex scenes, like (a) Crowded Street, (b) Highway, (c) Suburb, (d) Countryside, (e) Indoor, (f) Forest, (g) Desert, (h) Conventional Street, (i) Natural Park, (j) Rocky Terrain, (k) Real-world off-road, covering most common road conditions that HADAR may find applications in. <p>We have tested TeX vision, detection and ranging, and reported in this revised manuscript the (i) real-world HADAR performance and (ii) generalized performance on dissimilar scenes. We are glad to confirm that HADAR has promising and consistent performance that beat AI-enhanced thermal sensing. See Fig.1c, Fig.6, Extended Data Figs.2-8 in the revised manuscript for more details. Please see the supplemental video for a demonstration of real-world TeX vision of an off-road scene at night!</p>

We have also made revisions according to all other comments. Now, we will address each comment sequentially in the following. Notations used in this response include C: Comment, R: Reply, *Italic*: revisions, underline: emphasize.

This permanent link to the HADAR database (<https://github.com/FanglinBao/HADAR>) will be made public once the paper is published. The temporary Microsoft one drive link for reviewers is:

https://purdue0-my.sharepoint.com/:f/g/personal/baof_purdue_edu/ErtlrHN6qO1IvNtfbD9ezaIBDPtdSjldpW7EEegMuPw_RQ?e=MzbG6V

Reviewer 1	
C0	<p>The reviewer recognizes and appreciates the diligence and effort shown by the authors in addressing the provided comments and concerns. While the authors have done a great job in laying out the theoretical framework and have provided sufficient mathematical and physics-based explanations for their approach, some major issues remain concerning the scientific soundness of the proposed ML/DL approaches and their correctness. The work certainly has the potential to present a breakthrough in the field of computer vision and AI. However, the majority of the testing is done on synthesized images but makes claims that it can outperform models that are largely applied to real-world imagery and the noise that such imagery inevitably encompasses. To justify these claims, the reviewer’s greatest concern is that the synthetic datasets used by the authors are not comparable to real-world datasets acquired by HADAR. Despite the novel hardware design, the authors pursue to majorly utilize the synthetically generated datasets which cripple the proposed novel design. To truly prove the claim of superior performance and provide the breakthrough alluded to in the paper, the reviewer needs to see similar performance when using real-world datasets as opposed to results-based majorly on synthetic datasets.</p>
R0	<p>We would like to thank the reviewer for the encouraging response and valuable comments. We have addressed each comment individually below and made major revisions to improve the quality of this manuscript. Overall, we are glad to confirm and report that HADAR has promising performance over multiple dissimilar scenes as well as real-world experimental scenes, consistently outperforming AI-enhanced thermal sensing.</p> <ul style="list-style-type: none"> • Explicitly, reply R1 reports the overall HADAR performance over multiple dissimilar scenes • R2 and R3 provide the details of the TeX-Net code, training, cross validation, and how TeX vision is generated • R4 and R6 report and analyze the real-world experimental HADAR performance and statistics • R5 discusses the human-robot identification with thermal people detection networks <p>We also want to point out that, in this revised manuscript, we have improved our argument about HADAR ranging to better reflect our findings.</p> <p><old argument>: HADAR ranging shows an accuracy improvement up to two orders of magnitude compared with existing thermal ranging.</p> <p><new argument>: <i>HADAR ranging at night beats existing thermal ranging and shows an accuracy comparable with RGB stereovision in daylight.</i></p> <p>In this revised manuscript, we shall present results accordingly to support the new argument. The reason we improve the argument is given as below.</p> <p>Note that RGB stereovision in daylight already has widespread applications in autonomous navigation, while thermal ranging at night (stereovision based on thermal IR images) is still elusive due to the ghosting effect. Since RGB camera records no signal in the dark, having night vision and ranging that are comparable to daylight counterparts has been a long-standing quest in machine perception. One of our themes throughout this manuscript is ‘HADAR and TeX vision see textures and depth through the darkness as if it were day’. It is most natural to use RGB stereovision in daylight as a baseline and quantitatively compare TeX and IR at night with RGB</p>

	<p>stereovision in daylight. We think the new argument is more suitable because it directly meets the need in applications and is scene-independent and universal.</p> <p>In contrast, the old argument involved the depth accuracy improvement of HADAR ranging vs. thermal ranging. As we have explained in the previous round of revision, this accuracy enhancement is scene dependent, metric dependent, and not universal. In this work, we have still observed in many scenes that the accuracy improvement is near 100X. But for some other simple scenes where thermal ranging itself can be comparable to RGB stereovision, HADAR ranging can match RGB stereovision but the accuracy enhancement over thermal ranging will not always be as high as 100X. Please see Fig.6 in the main text and Fig.S19 in the Supple. Info. for quantitative statistics.</p>
C1	<p>The following items present the remaining concerns of the reviewer which to be addressed before the article can be accepted:</p> <ol style="list-style-type: none"> 1) The reviewer acknowledges the work done to create the new spectro-spatial deep learning model presented in the paper compared against the multi-layer perceptron as presented in the earlier version of the manuscript. Still, there are some concerns about the generalized performance of the proposed model for multiple scenarios where the scenes depicted, vary as is typical of real-world imagery. So far in the paper, only select scenes have been presented where the HADAR model beats the SOTA models. It is equally important to gauge the model's performance over multiple images that are dissimilar and report the overall performance on these datasets.
R1	<p>We agree with the reviewer that it is equally important to gauge the model's performance over multiple dissimilar images and report the overall performance of HADAR. To make HADAR more convincing and go beyond selected scenes, in this revised manuscript we have made our best efforts to create a HADAR database consisting of 11 dissimilar datasets and test TeX vision, detection and ranging on it.</p> <ol style="list-style-type: none"> 1. <u>HADAR database:</u> The HADAR database includes dissimilar scenes like Crowded Street, Highway, Suburb, Countryside, Indoor, Forest, Desert, etc., covering most common road conditions that HADAR may find applications in. The 11th dataset is a real-world off-road scene with heat cube dimension Height*Width*Channel = 260*1500*49, while the first 10 scenes are synthetic with heat cube dimension Height*Width*Channel = 1080*1920*54. The channels in the real-world scene correspond to the 5th ~ 53rd channels of the synthetic scenes. The HADAR database is designed to mimic various dissimilar self-driving situations. The HADAR sensor(s) are either mounted at the positions of headlights, or on the top of the automated vehicles, or on robot helpers. Each scene has 5 frames for each camera, and there are 30 different kinds of materials in total in the HADAR database. For the Street, Suburb, Rocky Terrain, and the Real-World Off-Road scenes, TeX, RGB and IR images are provided for the purpose of ranging. The Street scene has a long animation version (100 frames, 12 channels). For the real-world experimental scene, HADAR sensor is a pushbroom hyperspectral imager that can produce 256 spectral bands (price ~ \$1M). Sensor parameters are given in Methods-prototype HADAR calibration. Data collection was conducted under the DARPA IH (The Defense Advanced Research Projects Agency, Invisible Headlights) project, with the help of the Army night-vision team (Infrared Camera Technology Branch, DEVCOM C5ISR Center,

U.S. Army). Ground truth material library was not collected. Instead, we used a customized TES (temperature-emissivity separation) algorithm followed by K-means clustering to generate a semantic library as an estimation of the material library. 4 frames that share the same semantic library form the 11th dataset. The heat cubes have been interpolated into 49 channels to match the channels in synthetic scenes developed before the experiments. Only 49 channels of all the scenes are used to train TeX-Net.

The HADAR database is summarized in Extended Data Figs.2-3, as cited below. The database is available at <https://github.com/FanglinBao/HADAR> in the HADAR database folder.

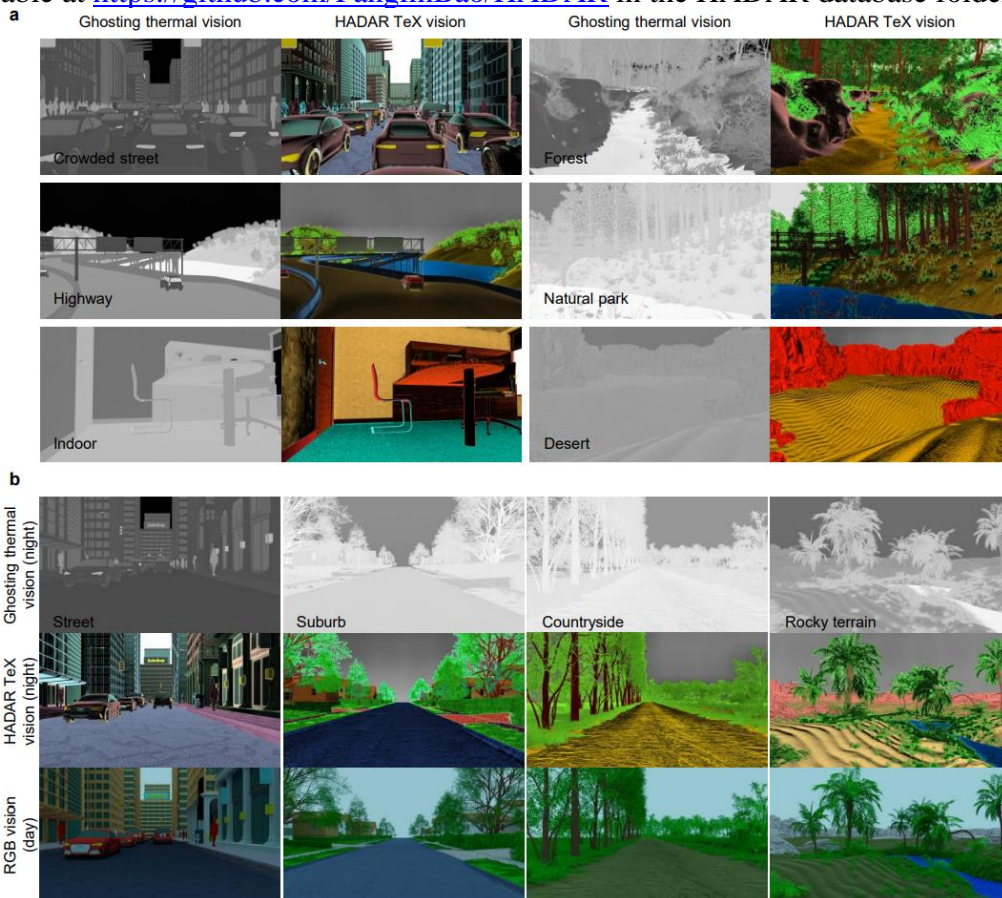




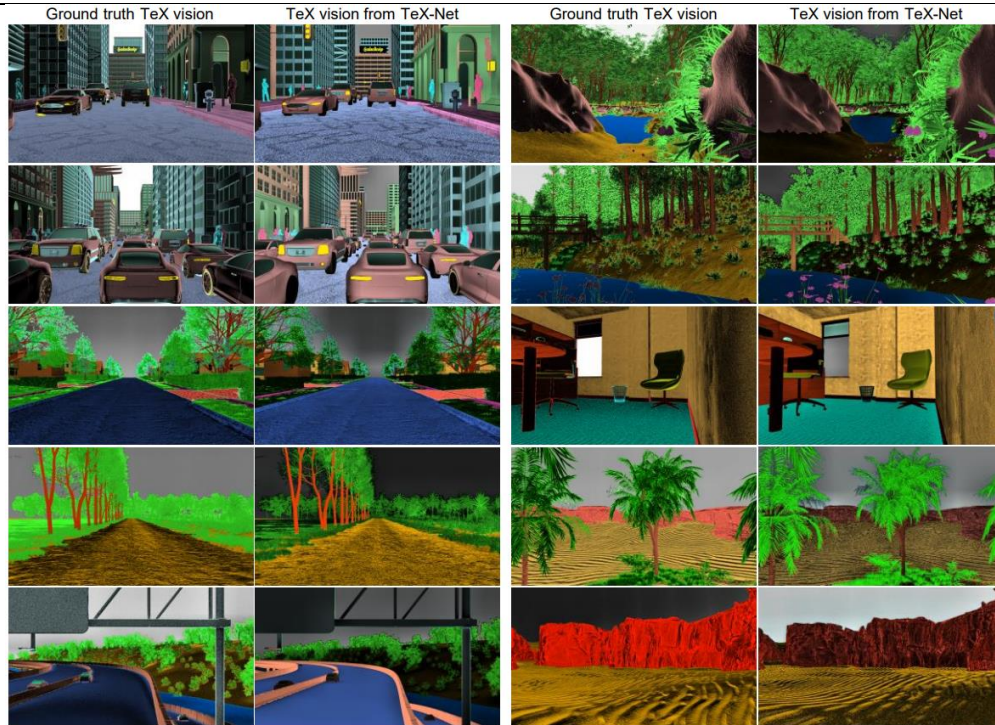
Fig.1 HADAR database with 11 dissimilar scenes to show the generalized performance of HADAR.

2. TeX-Net and TeX vision:

We split the HADAR database (11 scenes) into training set + validation set to train the TeX-Net with 5-fold cross validation. Due to limited experimental data, we manually split the database, instead of randomly splitting, to ensure the same diversity of the validation set and training set. We used a hybrid loss with half supervised loss and half physics loss, and we trained TeX-Net for 40K epochs. Since the real-world scene (260*1500) has a different image size with the synthetic scenes (1080*1920), we used random crop (256*256) in training.

For the experimental scene, we first applied our newly proposed TeX-SGD (semi-global decomposition) to generate the TeX vision, as an estimation of the ground truth TeX vision. TeX-SGD results are then used together with synthetic data to train the TeX-Net. TeX-SGD is a non-machine-learning approach that decomposes TeX pixel per pixel based on the physics loss and a smoothness constraint.

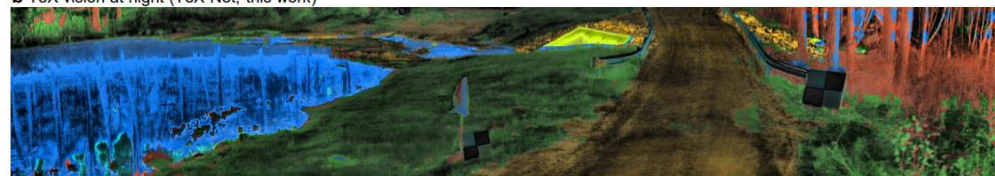
The TeX-Net performance (TeX vision) on the validation set has been summarized in Extended Data Fig.3 and Fig.S18 in the Supple. Info., as cited below. The network model, training details, and pre-trained weights are available along with the database at <https://github.com/FanglinBao/HADAR> in the TeX-Net folder. Codes for TeX-SGD and TeX visualization are also available there in the HADAR folder.



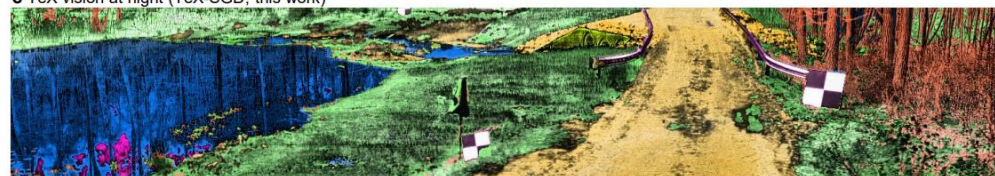
a Ghosting thermal vision at night (real world)



b TeX vision at night (TeX-Net, this work)



c TeX vision at night (TeX-SGD, this work)



d RGB vision in daylight



Fig.2 General TeX vision performance on various dissimilar scenes.

Interestingly, for the real-world scene, even though the TeX vision generated by TeX-SGD was used as the ground truth, TeX-Net outputs a spatially smoother TeX vision than TeX-SGD. This is because TeX-Net utilizes both spatial and spectral information in TeX decomposition, while TeX-SGD only uses the spectral information. We observed that TeX-SGD is better at material identification and texture recovery for fine structures, such as, bridge fence, bark wrinkles, culverts, etc.

3. Ranging:

As mentioned in Reply R0, we have improved our ranging argument to ‘HADAR ranging at night beats existing thermal ranging and shows an accuracy comparable with RGB stereovision in daylight’. Therefore, we have done ranging statistics accordingly on various dissimilar scenes and real-world scenes to support the new claim. The performance figure is given in Fig.6 and Fig.S19, as cited below. Fig. c shows that HADAR ranging metrics beat thermal ranging and are comparable to RGB stereovision, clearly demonstrating our new argument.

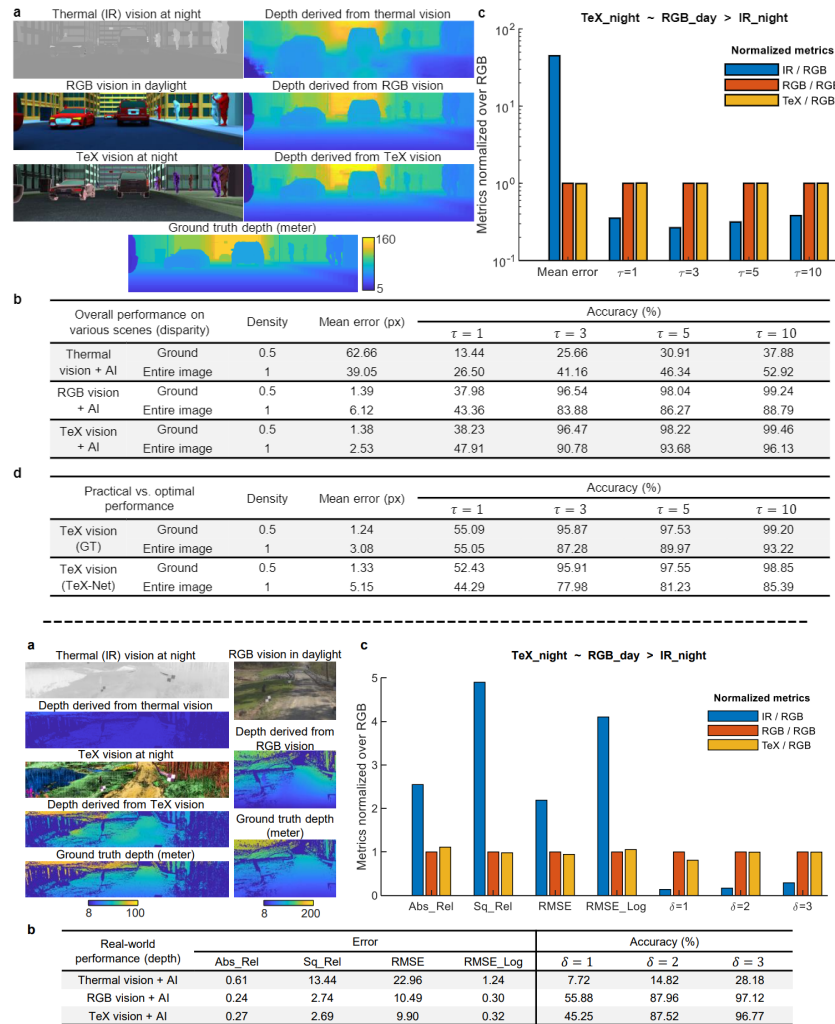


Fig.3 General and real-world HADAR ranging performance.

The data for ranging performance on various dissimilar scenes and real-world scenes is available in the HADAR ranging folder in the HADAR database. We emphasize again that with different AI algorithms, with different metrics, or with different scenes the depth accuracy enhancement of TeX vs. IR may be different, but our argument about the depth accuracy relation ‘TeX_night ~ RGB_day > IR_night’ is robust, i.e., HADAR sees depth through the darkness as if it were day and beats thermal ranging.

4. Segmentation:

Our segmentation statistics on both synthetic and real-world scenes is given in Extended Data Fig.8, as cited below. Note that existing segmentation AI algorithms like DANet are usually trained on city scenes (e.g., CityScapes dataset). It is expected that they would have poor performance on non-city scenes beyond their training set. Therefore, to define a fair comparison, we only compare HADAR vs. AI-enhanced thermal sensing for City scenes in the HADAR database, see Fig.a-d and the upper table in the following cited figure.

Other scenes of the HADAR database are designed as dissimilar non-city scenes for the purpose of diversity. We have shown one typical example of them, the real-world off-road scene in Fig.e-h and the lower table. Note that the TeX vision here is generated by TeX-SGD (non-machine-learning approach), and we use a non-machine-learning algorithm to convert material map to semantics as well. As anticipated, DANet + thermal vision has a poor segmentation performance. In contrast, HADAR has achieved segmentations with mIoU>65%. This off-road statistic clearly shows the robustness of HADAR semantics on scenes beyond the training set, as HADAR semantics here is purely physics-driven.

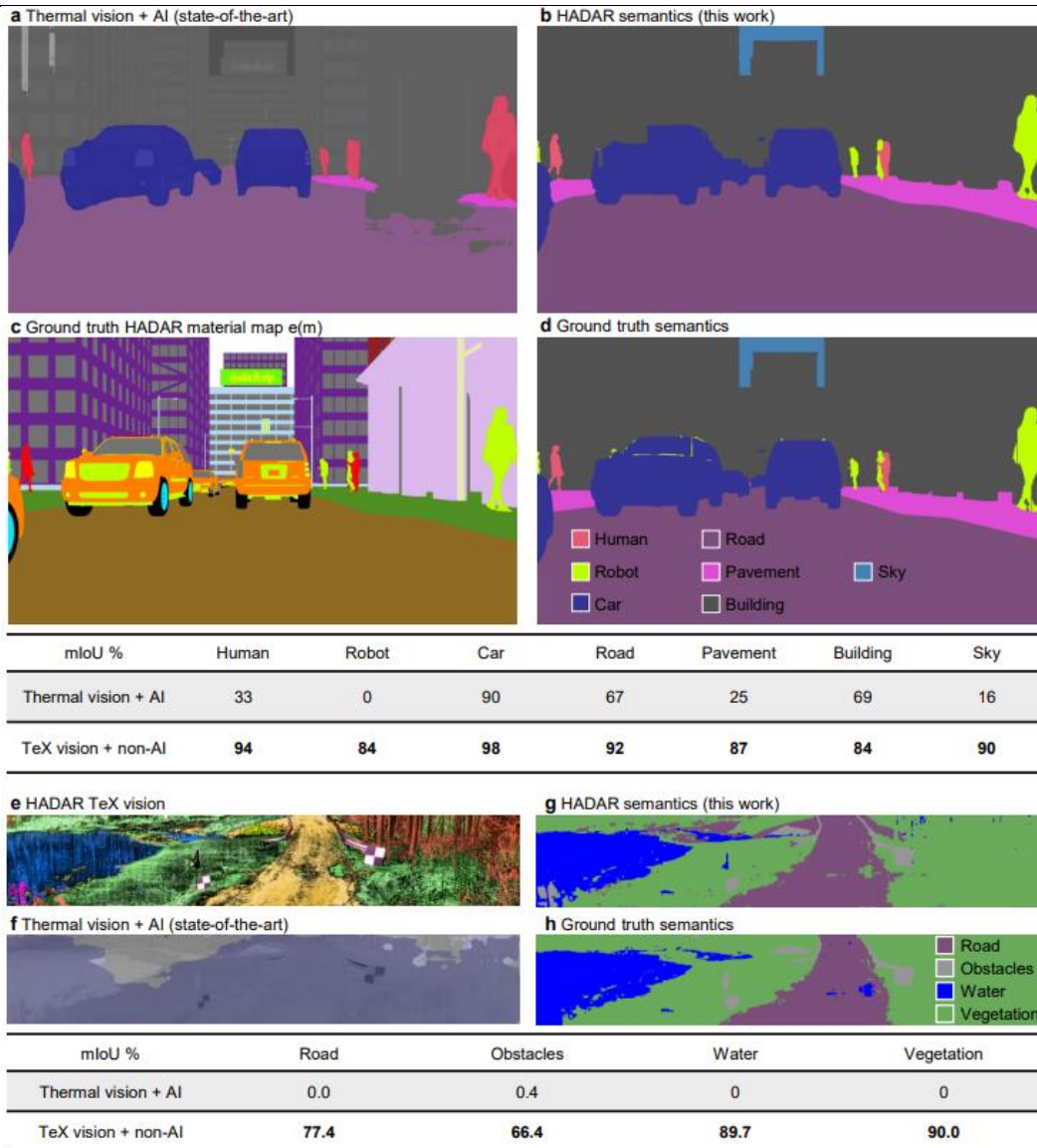
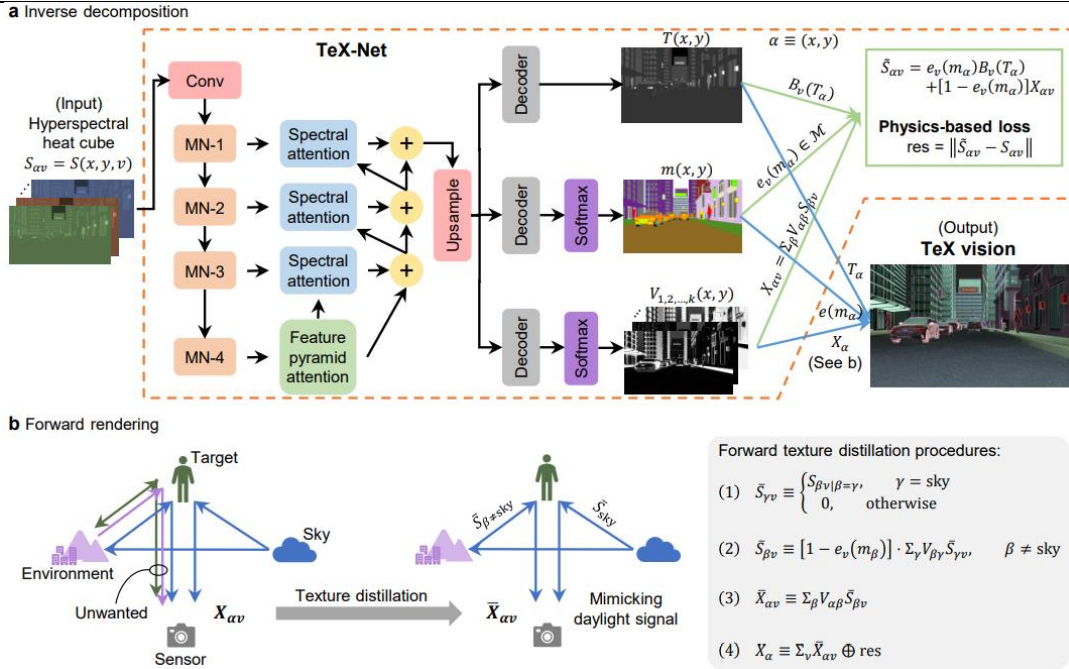


Fig.4 General HADAR segmentation performance.

C2 2) The reviewer is keen to have a look at the TexNet code to examine the parameters and their efficacy. It is not clear if the performance reported in the paper is for one specific test set. The inclusion of cross-fold validation results on the dataset would eliminate the bias in test performance.

R2 We thank the reviewer for pointing out the necessity of using cross validation. In this revised manuscript, we have used 5-fold cross validation in training TeX-Net and for all statistics. Explicitly, the HADAR database (11 scenes) was split into training set (80% data) + validation set (20% data). In each fold, one frame per view of each scene was selected for validation. This is to guarantee a similar diversity of the validation set as the training set. The TeX-Net codes, pre-

	<p>trained weights and loss curves are available at https://github.com/FanglinBao/HADAR in the TeX-Net folder.</p>
C3	<p>3) It is not clear how the authors generate TeX vision from T, E, X components. Is the TexNet framework trained end to end? Specifically, since the TexNet is trained with a physics-based loss on a different objective, how are the authors able to realize Tex vision out of the learned model? It is very important to elaborate these details to the readers as this is one of the fundamental strengths of this paper.</p>
R3	<p>We regret that it was not made clear as to how we generate the TeX vision from TeX components. TeX vision is a physics-driven representation of heat signals that takes into account Kirchhoff's laws as well as Planck's law of blackbody radiation. TeX vision gives a color image where the hue, saturation and value are physical attributes of temperature (T), emissivity (e) and texture (X). We do not train TeX-Net end to end due to TeX degeneracy as differing physical attributes of T, e, X can lead to the same observed spectrum. This is explained in the caption of Extended Data Fig.1. Instead, thermal light factor V has to be learnt by the model and is used to construct the texture X. In this revised manuscript,</p> <ol style="list-style-type: none"> <u>we have explicitly illustrated in Extended Data Fig.1b how X is obtained from V and physics loss function called 'res'</u> <ol style="list-style-type: none"> We first use the solved thermal lighting factors V and the sky illumination to construct the scattering signal (\bar{X}) mimicking daylight imaging signal. We call this process 'texture distillation' as it removes some unwanted scattering signal not originated from the sky. Please note that we have proposed both TeX-SGD (semi-global decomposition, non-machine-learning approach) and TeX-Net (machine learning approach) for TeX decomposition. Even though TeX-Net and TeX-SGD are minimizing the physics-based loss called 'res', some ground truth texture still remain in the physics-based loss 'res', due to cutoffs on scattering and number of environmental objects in our heat signal model in solving the inverse problem. Therefore, the final texture X is an information fusion of the distilled texture \bar{X} and the physics-based loss 'res'. The fusion process can be an image superposition or image fusion, as we have implemented in the code package. In cases where sky illumination is not specified, 'res' gives the final texture X. Advanced fusion deserves future exploration. <u>we have added a detailed paragraph in Sec. SIIC of the Supple. Info. to explain the procedures of generating TeX vision from T/e/X components</u> <u>we have provided our implementation code package (matlab) with sample data at https://github.com/FanglinBao/HADAR. See TeX.distillX in the code package for the texture distillation process (getting X) and TeX.Vision for the visualization process (getting TeX vision).</u> <p><u>All these details have also been added in Sec. SIIC of the Supple. Info. In the following, we cite our revisions for your convenience.</u></p> <p>Revision 1: new Fig.1b</p>



Revision 2: new figure caption (b) of Fig.1

(b), Texture distillation reconstructs the part of scattered signal that is originated only from sky illuminations. The texture distillation process is to mimic daylight signal as to form TeX vision, and it is done by evaluating the HADAR constitutive equation in a forward way, with the physical attributes solved out in TeX-SGD or TeX-Net. It removes the unwanted effect of other environmental objects being the light source which is unfamiliar in daily experience. The process can be described in 4 steps. Here, step (1) is the initialization that keeps only the sky illumination on and turns other radiations off. Step (2) is the iterative HADAR constitutive equation without direct emission. Evaluating it multiple times gives the multiple scattering effect. Note that the ground truth texture partly remains in the physics-based loss, res, due to cutoffs on scattering and/or number of environmental objects. The final estimated texture in step (4) is a fusion of distilled scattered signal \bar{X}_{av} and the physics-based loss res. Arrows in (b) indicate thermal radiation emitted/scattered along the arrow direction. The TeX-Net code, pre-trained weights, and a sample implementation of texture distillation is available at <https://github.com/FanglinBao/HADAR>.

Revision 3: how to generate TeX vision

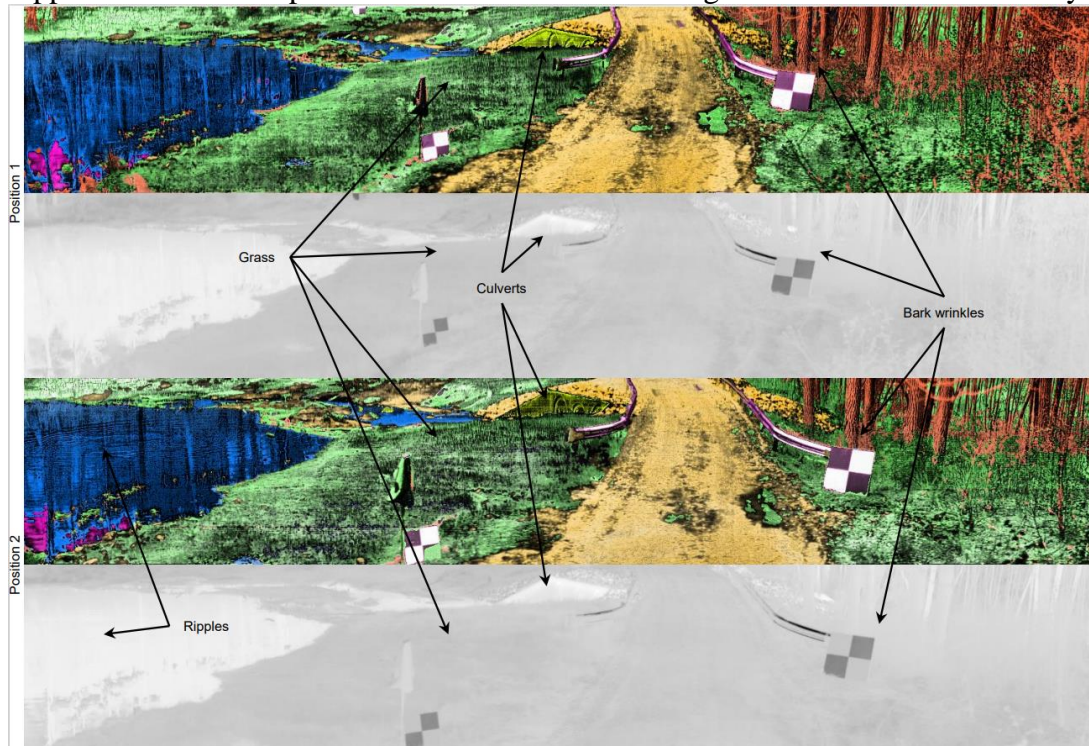
Explicitly, here we show how a TeX vision image is formed from the $T(x, y)$, $e[m(x, y)]$, and $X(x, y)$ output by TeX decomposition (through TeX-Net or SGD). (1), The material library \mathcal{M} comes with a hue library, \mathcal{H} . For each material in \mathcal{M} , we've assigned to it a hue value corresponding to its typical color as can be seen in daylight, so that the TeX vision will be similar to the familiar RGB image. For example, a hue value 90/255 corresponding to 'Green' is assigned to the material 'Vegetation', a hue value 160/255 corresponding to 'Blue' is assigned to the material 'Water', and a hue value 40/255 corresponding to 'Yellow' is assigned to the material 'Sand'. The hue channel of the TeX vision image is given by the following matlab pseudo code, $H = \text{reshape}(\mathcal{H}(m), \text{size}(m))$. (2), The saturation channel of the TeX vision image is given by the following matlab pseudo code, $S = \text{rescale}(T, 0, 1, \text{'InputMax'}, \text{tMax}, \text{'InputMin'}, \text{tMin})$, with user customized temperature range. (3), The Brightness channel of the TeX vision image is given by the following matlab pseudo code, $V = \text{adaphisteq}(\text{rescale}(X, 0, 1))$. (4), Finally, the TeX vision image is given by a color space transform, $\text{texIMG} = \text{hsv2rgb}(\text{cat}(3, H, S, V))$. Sample data and codes are available along with the HADAR database at <https://github.com/FanglinBao/HADAR>. Fig. S20 further shows an example of TeX vision image, in comparison with thermal vision and RGB vision.

C4	<p>4) The authors are presenting their HADAR framework as a real-world breakthrough as opposed to a theoretical design and attempt to demonstrate real-world efficacy. However, the results presented are largely based on synthetic images. For example, Figures 1, 2,3,5,6,7,8, &11 are all based on synthetic images and correspond to results tables. In contrast, Figures 4, 10, & 12 are from a real-world image but it is the same image and none of these figures correspond to results tables showing model efficacy . Thus, the reporting on comprehensive model performance is centered on images that do not accurately represent the noise found in real-world data and thus can not be representative of real-world efficacy, only validation of theoretical concepts. In other words, there exists a huge domain gap between the presented synthetic LWIR images and the HADAR captured images. While the synthetic images are valuable for some demonstration, the reviewer suspects model performance was gauged solely on synthetic images and thus the results presented do not reflect real-world performance. This is a major issue with the paper.</p>												
R4	<p>We agree with the reviewer that it is necessary to show real-world statistics of HADAR performance to claim the advantage of HADAR vs. AI-enhanced thermal sensing. In this revised manuscript, we have made our best efforts to include the HADAR prototype-II experiments and real-world performance statistics. We are glad to report that HADAR does show consistent and promising real-world performance that beats state-of-the-art AI-enhanced thermal sensing.</p> <p>1. <u>Sensor and experimental details:</u></p> <p>The HADAR prototype-II sensor is a pushbroom hyperspectral imager that can produce 256 spectral bands (price ~ \$1M). Sensor parameters have been detailed in Methods—prototype HADAR calibration. Data collection was conducted under the DARPA IH (The Defense Advanced Research Projects Agency, Invisible Headlights) project. There are multiple practical challenges in experiments, such as, (1) the pushbroom sensor shows horizontal streak noise due to dynamic drift of pixel gain and offset, (2) ground truth material library was not collected, and (3) the sky, which is a significant environmental object, was not directly observed. We have added one section (Sec. SV) in the Supple. Info. to explain the details of denoising, LiDAR-HADAR extrinsic calibration, and estimating the material library as well as the sky radiance.</p> <table data-bbox="568 1281 1169 1407"> <tr> <td>SV. HADAR prototype-2: experiments</td> <td>69</td> </tr> <tr> <td> A. Denoise</td> <td>69</td> </tr> <tr> <td> B. Extrinsic calibration between LiDAR and imaging sensors</td> <td>69</td> </tr> <tr> <td> C. Semantic library estimation</td> <td>70</td> </tr> </table> <p style="text-align: center;">2</p> <hr style="width: 50%; margin: 20px auto;"/> <table data-bbox="568 1554 1169 1638"> <tr> <td>D. Texture comparison and analysis between TeX vision and RGB vision in experiments</td> <td>71</td> </tr> <tr> <td>E. TeX-RGB image fusion in comparison with IR-RGB image fusion</td> <td>72</td> </tr> </table> <p>The relevant experimental HADAR data has 17 frames, among which, 3 frames have corresponding high-resolution LiDAR data. All 17 frames of TeX vision generated by TeX-SGD (semi-global decomposition) are available in the HADAR ranging folder in HADAR database given before. Note, we only use 4 frames for training TeX-Net. Since the pushbroom sensor was used along with multiple other sensors (irrelevant to this work) in the IH project, the data collection took so long that we observed significant changes of the estimated sky</p>	SV. HADAR prototype-2: experiments	69	A. Denoise	69	B. Extrinsic calibration between LiDAR and imaging sensors	69	C. Semantic library estimation	70	D. Texture comparison and analysis between TeX vision and RGB vision in experiments	71	E. TeX-RGB image fusion in comparison with IR-RGB image fusion	72
SV. HADAR prototype-2: experiments	69												
A. Denoise	69												
B. Extrinsic calibration between LiDAR and imaging sensors	69												
C. Semantic library estimation	70												
D. Texture comparison and analysis between TeX vision and RGB vision in experiments	71												
E. TeX-RGB image fusion in comparison with IR-RGB image fusion	72												

radiance throughout the experiment. The inaccurate sky radiance estimation causes performance fluctuations of TeX vision, as can be observed in 17 frames of TeX vision images. It also leads to instability of the estimation of the material library. Therefore, only 4 frames that share the same semantic library (estimated material library) are given in Scene-11 in the HADAR database to train the TeX-Net. Note that TeX-Net requires an input of material library in training. In the future, all these practical restrictions can be relieved with a proper on-site experimental characterization of the sky radiance and the material library.

2. TeX vision and texture recovery in real-world scenes:

We have proposed in this revised manuscript both non-machine-learning and machine-learning algorithms for TeX vision. The former is the TeX-SGD (semi-global decomposition), and the latter is the TeX-Net. TeX-SGD uses the spectral information to solve TeX decomposition pixel per pixel, based on the physics loss and a smoothness constraint, while TeX-Net utilizes both spatial and spectral information. We used the TeX vision generated by TeX-SGD to estimate the ground truth TeX vision and train TeX-Net. We observed that TeX-Net outputs spatially smoother TeX vision than TeX-SGD, with the help of spatial information. Currently, we observed that TeX-SGD is better at material identification and texture recovery for fine structures, such as, bridge fence, bark wrinkles, culverts, etc. Extended Data Fig.4 shows the material identification and texture recovery in comparison with traditional thermal vision for two sample frames. Extended Data Fig.3 shows the comparison of TeX vision from TeX-SGD vs. the TeX-Net. Both are cited as below. Also see Fig.S33 in the Supple. Info. for a comparison between TeX vision at night and the RGB vision in daylight.



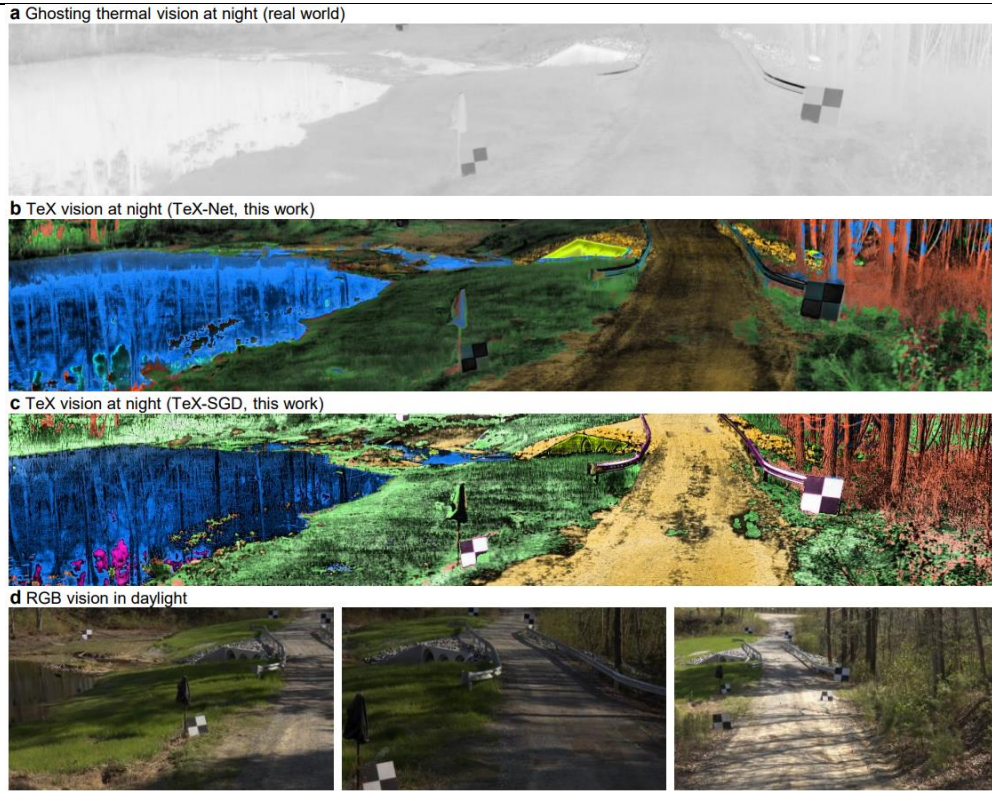


Fig.5 TeX vision and texture recovery for real-world scenes at night (completely dark).

3. HADAR ranging in real-world scenes:

The real-world HADAR ranging performance and statistics is given in Fig.6, as cited below. Ground truth depth is from the high-resolution LiDAR data. Estimated depth is visualized in the region where there is LiDAR data. Statistics is also done wherever there is LiDAR data. The experimental result Fig.(c) below is a visualization of the table (b). The metrics clearly demonstrate that HADAR ranging at night beats thermal ranging and matches RGB stereovision in daylight, and hence it demonstrates our argument ‘HADAR sees texture and depth through the darkness as if it were day’.

Due to practical restrictions in experiments as explained in point 1 (i.e., material library and sky radiance were not collected in experiments; sky radiance varied throughout the experiment), only 4 frames sharing the same estimated material library are used in TeX-Net, while the total TeX vision frames in TeX-SGD are 17. Therefore, we used TeX-SGD results to test real-world HADAR performance. Since the IH government experiment only collected monocular data, we used monocular depth estimation for HADAR ranging. We emphasize that our HADAR theory applies to both monocular and binocular stereovision (see Fig.S19 of the Supple. Info. for binocular stereovision). In the future, our work can be extended to binocular stereovision with this advanced sensor.

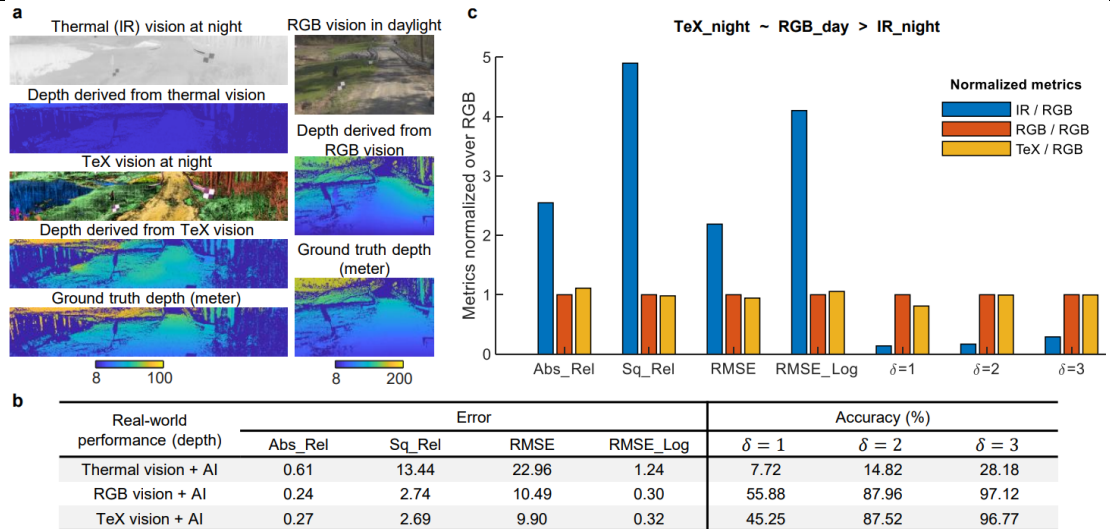


Fig.6 Real-world HADAR ranging performance at night. HADAR ranging at night beats thermal ranging and is comparable with RGB stereovision in daylight.

4. HADAR semantics in real-world scenes:

Our real-world HADAR semantics is given in Extended Data Fig.8, as cited below. As existing AI algorithms for segmentation are usually trained for city scenes (e.g., DANet trained on CityScapes), they are expected to have poor performance on off-road scenes. In contrast, HADAR semantics is physics driven. Both TeX-SGD and the conversion algorithm from material map to semantic map are non-machine-learning approaches. Therefore, it is expected to see that HADAR definitely has a major advantage for robust segmentation on scenes beyond the training set.

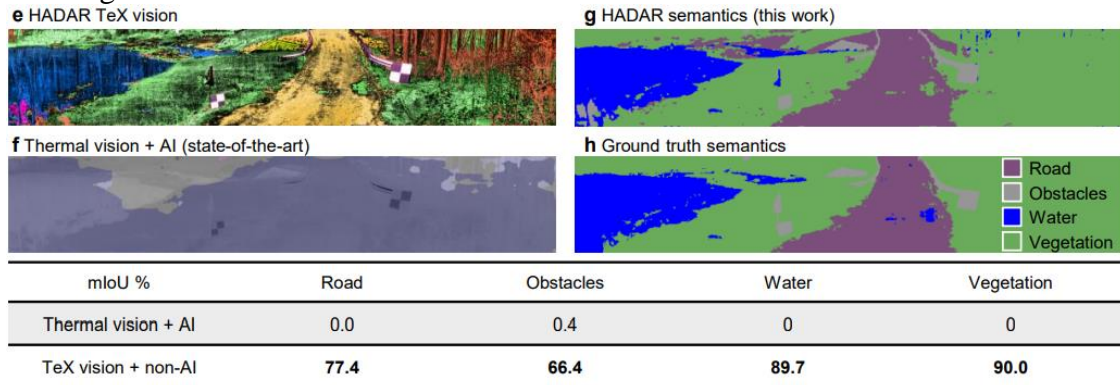


Fig.7 Real-world HADAR segmentation performance at night. HADAR semantics outperforms AI-enhanced thermal semantics.

C5) 5) Although the authors tested the figures with the RGB image pre-trained people detection framework, the domain gap between source and target image domains would lead to incorrect detection for target images. Be careful that although the framework correctly identifies humans for this specific frame, a generalized performance estimation over multiple frames would be useful. Still, it is recommended to train the detection model with target images with few-shot learning to avoid the requirement for large labeled datasets. Also, be informed that the thermal

	<p>and the HADAR results don't have significant texture information to correctly estimate the HOG features to be able to perform detection on these features. This is also evident with the detection score as shown in extended data fig5. Alternately, you could use SOTA human detection frameworks designed for thermal images. Take into consideration that the detection framework suffers if the HOG features generated for both the Robot and humans are the same since the only difference between these is the semantic color and some textures.</p>
<p>R5</p>	<p>We thank the reviewer for the suggestion of using a detection model designed for thermal images. In this revised manuscript, we have adopted a very recent (from 2022) thermal-YOLO model fine-tuned on thermal automotive dataset, see Extended Data Fig.7. Also, we have tested with the people detection model in standard computer vision toolbox in matlab R2021b, in addition to our previous test in matlab R2018b. They all show consistent results. The thermal-YOLO result was shown in the manuscript. The human vs. robot detection result is cited as below. Note that HADAR detection here was performed on corresponding material region (using the physical attributes). The advantage of HADAR detection clearly comes from TeX vision and is independent of used AI algorithms.</p> <div data-bbox="435 846 1256 1478" data-label="Figure"> </div> <p>Fig.8 HADAR detection beats AI-enhanced thermal detection.</p>
<p>C6</p>	<p>6) As the authors pointed out, the identification of the object is dependent on the semantic/statistical distance of each material in the library, which decreases when materials are introduced to the library. To address this issue, the authors propose to use a setup with higher multispectral band capabilities but their hardware itself is constrained in that it cannot identify even a minimal number of bands that occur in a real-world application. Firstly, these sorts of hardware constraints are not considered in synthetic examples which are used to demonstrate the superior performance on various tasks compared to thermal or other modalities. Secondly, realizing these constraints, I am skeptical of their performance in real-world situations where</p>

	<p>design costs are the bottleneck. The authors emphasize the need for a low-cost design for HADAR optimal design, but this will not allow the design to function optimally for optimal setup for separability of materials in the real-world situation. In the synthetic examples to prove HADAR efficacy, these constraints are relaxed as a result of which the authors can demonstrate many sets of materials whereas, for the real world, the material library is a bare minimum and insufficient to capture the variation in spectral signatures encountered in the real world.</p> <p>Please carefully address these concerns for the next submission.</p>
R6	<p>We agree with the reviewer that our previous prototype HADAR (prototype-I, low-end HADAR, the sensor + filters are about \$20K) was a proof-of-concept demonstration and not ready for complicated real-world scenes. In this revised manuscript, we have made our best efforts to experimentally demonstrate HADAR prototype-II (high-end HADAR, advanced pushbroom sensor cost ~ \$1M) in complicated off-road scenes. HADAR prototype-II can generate 256 spectral bands, and as we have demonstrated in experiments shown before, it can distinguish materials properly in real-world night scenes.</p> <p>We note that our work will motivate cost-effective CMOS-compatible thermal sensors in the future. Furthermore, research communities in coded apertures and metasurfaces will be motivated to develop high-speed, low-cost, low-complexity spectral thermal imagers as alternatives to traditional pushbroom cameras. One example is the CRISP architecture [1], which is a low-cost microbolometer platform with a coded aperture that can give more than 50 spectral bands in the long-wave infrared. We therefore believe the cost barrier to HADAR can be broken in the near-term.</p> <p>We would like to further point out that, even though our HADAR theory is explained in the form of explicit spectral resolution, the TeX vision isn't fundamentally restricted to the input of explicit spectral radiance, and therefore, spectrum reconstruction is not always essential. For example, our least-squares estimator is proposed for low-end HADAR applications where the 3rd axis of the heat cube is not wavenumber but filter index, and as we have demonstrated previously, it can distinguish certain amounts of materials. If one wants to distinguish an increasing number of materials, more filters are needed, and in principle, the number of filters is not limited. In Ref. [2], 195 filters have been used in experiments. In the future, we can replace the filter-wheel approach to Bayer filter mosaic approach on FLIR cameras, or otherwise, use customized filter holders that can support desired number of filters. With more filters, spectrum reconstruction will be increasingly more accurate, but again, it is not essential to convert the 3rd axis of heat cube from filter index to wavenumber. Furthermore, TeX-Net can also be trained in a way that the input heat cube has filter index rather than wavenumber.</p> <p>Two prototypes we have demonstrated in this work clearly show the functionality-cost balance of HADAR. We agree that in situations where design costs are the bottleneck, low-end HADAR may not be able to function optimally as the high-end HADAR does. However, this is a common trade-off that can also be seen in other sensors like LiDAR. For example, the low-end sparse LiDAR used in our work (Fig.5) cannot detect the black car 10 m away from the sensor. In a fair</p>

comparison between low-end HADAR and low-end LiDAR, as we have demonstrated, our low-end HADAR has better detection performance.

Accordingly, we have revised our manuscript to make our argument clearer.

<Methods -- prototype HADAR calibration>:

“In our proof-of-concept experiments, we used the filter-wheel approach to demonstrate the HADAR prototype-1. The filter-wheel approach is time consuming but cost effective, *suitable for low-end HADAR applications. In contrast, HADAR prototype-2 with a pushbroom sensor was demonstrated for high-end HADAR applications.* HADAR can also be implemented by other approaches with mosaic sensors, gratings, prisms, interferometers, or Fabry-Perot cavities, depending on the desired spectral resolution, spatial resolution, data acquisition speed, or functionality-cost balance.”

<Sec. SIVD of the Supple. Info.>:

“*First, we emphasize that spectrum reconstruction is not essential for TeX vision nor HADAR. It is useful when the explicit spectral resolution of radiance is desired, e.g., to help estimate the material library or environmental radiance in real-world experiments.* When sufficient filters are available, reconstruction ...”

Reference(s):

- [1] R. M. Sullenberger, A. B. Milstein, Y. Rachlin, S. Kaushik, and C. M. Wynn, "Computational reconfigurable imaging spectrometer," Opt. Express 25, 31960-31969 (2017)
- [2] Bao, Jie, and Mounqi G. Bawendi. "A colloidal quantum dot spectrometer." Nature 523.7558 (2015): 67-70.

We thank the reviewer once again for the time and efforts spent to provide us such helpful and detailed comments. With the above changes, we believe the manuscript is now significantly improved and ready for publication.

Cover letter to Reviewer 3

We would like to thank the reviewer for the encouraging response and valuable comments. It certainly helped us to significantly revise our manuscript. In this round of review, there is one remaining major concern from all reviewers. Here, we briefly list all the major revisions, and we will provide individual replies to each comment from the next page onwards.

Major concern: Will HADAR work in a real-world environment?

The reviewers want to see (i) real-world HADAR performance and (ii) generalized performance over multiple dissimilar scenes. Explicitly, the reviewers want to see the performance in more complex synthetic scenes such as dense crowds and dense vehicles, and in real-world open environments with inevitable noise and unknown classes of materials.

Major revision(s): Real-world demonstration of HADAR

In this latest revised manuscript, we have made major revisions to fully address the concerns.

- We have added HADAR prototype-2 and real-world experiments**, in the presence of sensor noise and unknown materials. Based on the input from the reviewer and to respond to the detailed questions, we formed a partnership with DARPA (The Defense Advanced Research Projects Agency, through the Invisible Headlights project) and the Army night-vision team (Infrared Camera Technology Branch, DEVCOM C5ISR Center, U.S. Army). We have now collected real-world experimental data using a pushbroom hyperspectral imager (~\$1M) and it took ~\$20K a day for personnel to collect data. Along with our previous home-built HADAR prototype-1 device, we call this as HADAR prototype-2. This data will be made available to the global research community accelerating progress not only in machine learning algorithms but also in the creation of new cost-effective and cheap sensors for HADAR. We have also generalized our HADAR theory so that it does not require an input of material library.
- We have created a HADAR database with 11 dissimilar scenes to test generalized HADAR performance.** Our HADAR database consists of complex scenes, like (a) Crowded Street, (b) Highway, (c) Suburb, (d) Countryside, (e) Indoor, (f) Forest, (g) Desert, (h) Conventional Street, (i) Natural Park, (j) Rocky Terrain, (k) Real-world off-road, covering most common road conditions that HADAR may find applications in.

We have tested TeX vision, detection and ranging, and reported in this revised manuscript the (i) real-world HADAR performance and (ii) generalized performance on dissimilar scenes. We are glad to confirm that HADAR has promising and consistent performance that beat AI-enhanced thermal sensing. See Fig.1c, Fig.6, Extended Data Figs.2-8 in the revised manuscript for more details. Please see the supplemental video for a demonstration of real-world TeX vision of an off-road scene at night!

We have also made revisions according to all other comments. Now, we will address each comment sequentially in the following. Notations used in this response include C: Comment, R: Reply, *Italic*: revisions, underline: emphasize.

This permanent link to the HADAR database (<https://github.com/FanglinBao/HADAR>) will be made public once the paper is published. The temporary Microsoft one drive link for reviewers is:

https://purdue0-my.sharepoint.com/:f/g/personal/baof_purdue_edu/ErtlrHN6qO1IvNtfbD9ezaIBDPtdSjldpW7EEegMuPw_RQ?e=MzbG6V

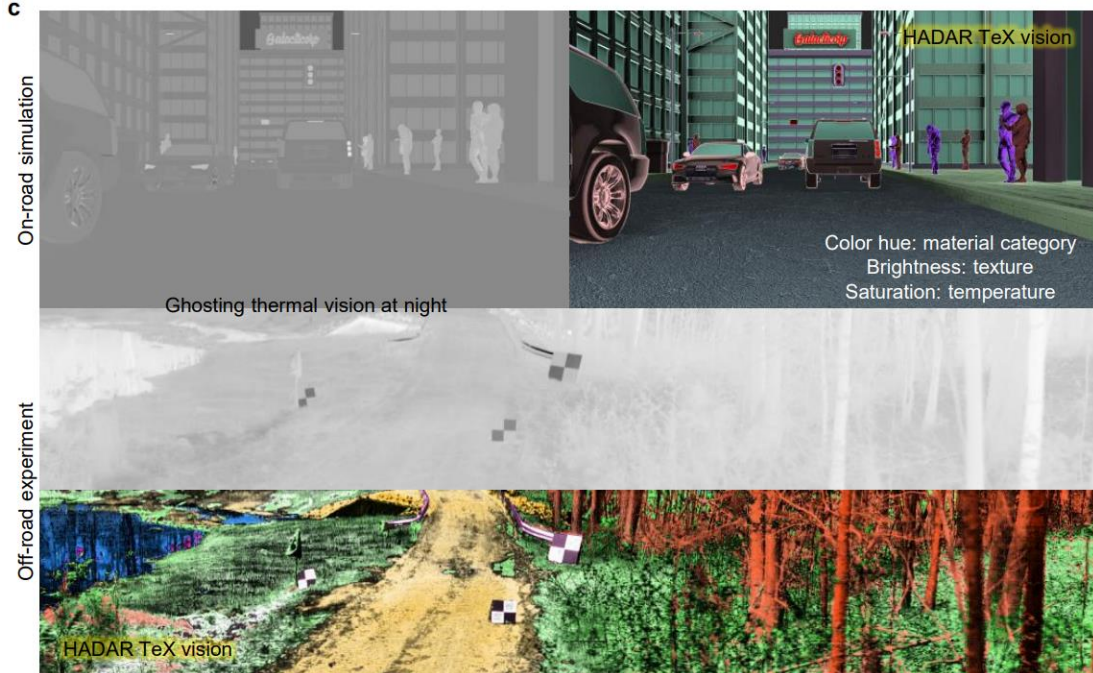
Reviewer 3	
C0	In this paper, the authors proposed and demonstrated HADAR (Heat-Assisted Detection and Ranging) for fully-passive and physically-aware machine perception. This work is interesting, the authors exploited physics-driven perception to achieve improved performance against AI-enhanced thermal sensing.
R0	We would like to thank the reviewer for the encouraging response and valuable comments. We have addressed each comment individually below and made major revisions to improve the quality of this manuscript.
C1	The paper is not easy to read, and I suggest authors should move some essential information from supplementary material to the article.
R1	<p>We regret that the previous version was not easy to read. During the review & revision process, the content of the manuscript became increasingly dense. We agree and we have made every effort to keep the essential information for the broad audience in the main text. We regret that it was not well done in the previous version. In this revised manuscript, we have even more results to show, particularly, the real-world HADAR experiments and the general HADAR performance on the HADAR database with 11 dissimilar scenes. To improve the readability, we have revised the contents' layout and added essential information in the main text.</p> <p><Old manuscript structure> The main text was devoted to describing the fundamental limits of HADAR. Real HADAR performance was not shown in the main text. HADAR thermography was an important subject but not closely related to autonomous navigation.</p> <p><New manuscript structure> <u>Since the key technical breakthrough of this work is overcoming the ghosting effect and beating thermal ranging, we have moved our new experimental results of real-world HADAR TeX vision (Fig.1c) and HADAR ranging (Fig.6) into the main text.</u> The figure for HADAR thermography is moved to Extended Data Fig.9 instead.</p> <p>Now, the revised main text firstly explains the origin of the ghosting effect and introduces the theory of TeX vision. Then, it illustrates the fundamental limits and real-world performance of HADAR detection and ranging, clearly demonstrating how HADAR addresses the existing challenges of phantom braking and thermal ranging. The main text uses condensed/concise scenes to illustrate the physics more clearly, leaving HADAR performance in general/complicated scenes in the Extended Data. The technical details of the HADAR theory and experiments are mainly given in the Supple. Info. This is to make the main text more friendly to the broad readership of Nature.</p> <p>Revisions: To make the layout of the contents clear in the main text so as to better guide the readers, <u>we have added explicit references to the Supple. Info. and Methods.</u> We have also added the following essential explanations in the main text.</p> <p><end of left column, page 3> '...With general HADAR performance shown in Extended Data, here we demonstrate the fundamental limits as well as real-world performance of HADAR.'</p>

<the whole section of *Real-world HADAR perception*, page 4>:

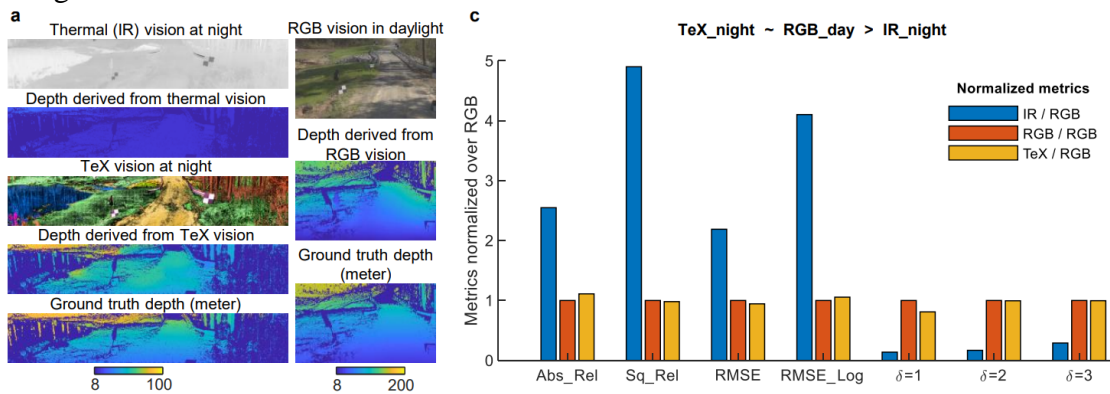
‘We now experimentally demonstrate HADAR in real-world scenes. *Our HADAR prototype-1 for low-end applications* is based on commercial FLIR thermal camera with custom designed spectral modules (see Extended Data Fig. 10) ...’

‘*Our HADAR prototype-2 for high-end applications is based on a pushbroom hyperspectral imager...*’


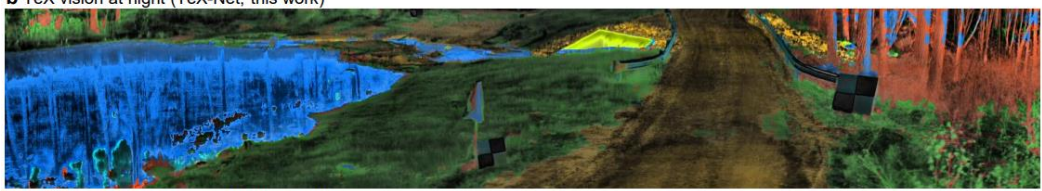
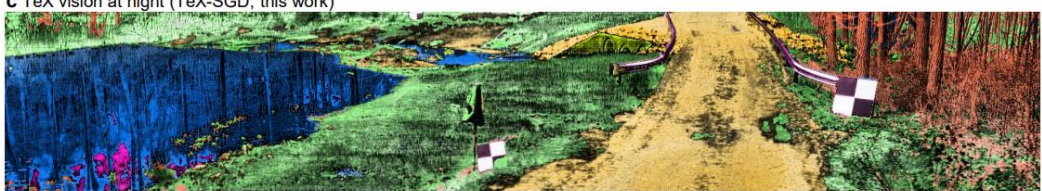

<new Fig.1c>:



<new Fig.6>



C2	<p>Following are a few suggestions and some questions for the authors:</p> <p>1. Although more complex data are used in experiments suggested by other reviewer, I still feel the scenes visualized in the article are not complex enough. I hope the authors will add more complex scenes such as dense crowds and dense vehicles in the additional content.</p>																																								
R2	<p>We thank the reviewer for the explicit suggestion. In this revised manuscript, we have created a HADAR database consisting of 11 dissimilar scenes. The HADAR database includes crowded scenes (e.g., Crowded Street), complex scenes (e.g., Forest), and real-world off-road scenes, with 30 different kinds of materials in total. A summary of the database is given in Extended Data Figs. 2 and 3. Our TeX vision, detection, and ranging performance is also based on these complex scenes, as shown in Fig. 6, Extended Data Figs. 3, 4, 7, and 8. The new complex scenes are cited below.</p> <div data-bbox="316 667 1404 1638"> <p>a</p> <table border="1"> <thead> <tr> <th></th> <th>Ghosting thermal vision</th> <th>HADAR TeX vision</th> <th>Ghosting thermal vision</th> <th>HADAR TeX vision</th> </tr> </thead> <tbody> <tr> <td>Crowded street</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Highway</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Indoor</td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table> <p>b</p> <table border="1"> <thead> <tr> <th></th> <th>Street</th> <th>Suburb</th> <th>Countryside</th> <th>Rocky terrain</th> </tr> </thead> <tbody> <tr> <td>Ghosting thermal vision (night)</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>HADAR TeX vision (night)</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>RGB vision (day)</td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table> </div>		Ghosting thermal vision	HADAR TeX vision	Ghosting thermal vision	HADAR TeX vision	Crowded street					Highway					Indoor						Street	Suburb	Countryside	Rocky terrain	Ghosting thermal vision (night)					HADAR TeX vision (night)					RGB vision (day)				
	Ghosting thermal vision	HADAR TeX vision	Ghosting thermal vision	HADAR TeX vision																																					
Crowded street																																									
Highway																																									
Indoor																																									
	Street	Suburb	Countryside	Rocky terrain																																					
Ghosting thermal vision (night)																																									
HADAR TeX vision (night)																																									
RGB vision (day)																																									

	<p>a Ghosting thermal vision at night (real world)</p>  <p>b TeX vision at night (TeX-Net, this work)</p>  <p>c TeX vision at night (TeX-SGD, this work)</p>  <p>d RGB vision in daylight</p>  <p>Fig.1 HADAR database with complex scenes.</p>
C3	<p>2.The authors mention that HADAR is expected to make great progress in industry, but it seems that there is no demonstration of HADAR computational efficiency and deploy-ability in the article.</p>
R3	<p>We thank the reviewer for pointing out the computational efficiency and deploy-ability of HADAR. In this revised manuscript, we have added a section in the Methods --- ‘Computational efficiency and deployability of HADAR’.</p>

	<p>Computational efficiency and deploy-ability (1) Our TeX-Net has about 0.5M weights in total. The evaluation of our TeX-Net (GPU Nvidia RTX A6000 48GB) takes 42.4 ms. Data collection of the currently used pushbroom hyperspectral imager takes around 1s, but the filter-wheel approach can be optimized down to around 10ms with high-speed filter wheel (<i>e.g.</i>, Telops multispectral cameras). Overall, our results show that HADAR data collection and processing can support up to 20 Hz TeX vision frame rate. Pursuing higher frame rate motivates further research on new hyperspectral imaging sensors to collect thermal infrared data and photonic neural networks for TeX decomposition. (2) Our generalized HADAR theory does not require the input of a material library and hence is free of on-site library collection/calibration. This enables real-time HADAR applications. Our HADAR prototype-2 experiment is a field test with the HADAR sensor mounted on a car. Corresponding TeX vision results on the DARPA IH test data shows the deploy-ability of HADAR, see Extended Data Figs. 3 and 4.</p>
C4	<p>3. In HADAR, a deep neural network approach is used to predict materials by spectra, and I'm rather curious about what the results will be for HADAR if the prediction is wrong in this process. Is the prediction of materials by spectra robust enough?</p>
R4	<p>We would like to answer the question in the following aspects and revise accordingly.</p> <ol style="list-style-type: none"> 1. <u>Robustness of material prediction:</u> According to our theory of HADAR identifiability (Eq. 2 in the main text, Algorithms 1-3 in the Supple. Info.), material prediction is quantified by the probability of correct prediction (detection probability). The detection probability as well as the HADAR identifiability depends on the collected photon number (signal strength), sensor noise, semantic distance and spectral resolution, and the explicit cutoff on scattering and environmental objects in solving TeX decomposition. Intuitively, the robustness of material prediction will increase (<i>i.e.</i>, detection probability increases) when the sensor noise decreases and/or the spectral resolution increases. 2. <u>Algorithms to improve material prediction:</u> In our TeX decomposition algorithms, TeX-Net utilizes the spatial information in addition to the spectral information to improve the robustness of material prediction. On the other hand, TeX-SGD (semi-global decomposition) uses a spatial smoothness constraint to improve the robustness of material prediction. 3. <u>Real HADAR performance when material prediction is wrong:</u> In case of wrong material prediction, temperature and thermal lighting factor estimation will be biased and the physics-based loss residue will be large. In the following, we discuss the consequences in TeX vision, detection, segmentation, and ranging, respectively. <ul style="list-style-type: none"> • TeX vision

TeX vision will have speckles within an otherwise uniform object. We have observed this in both TeX-Net and TeX-SGD results. See the following real-world experimental results for example.

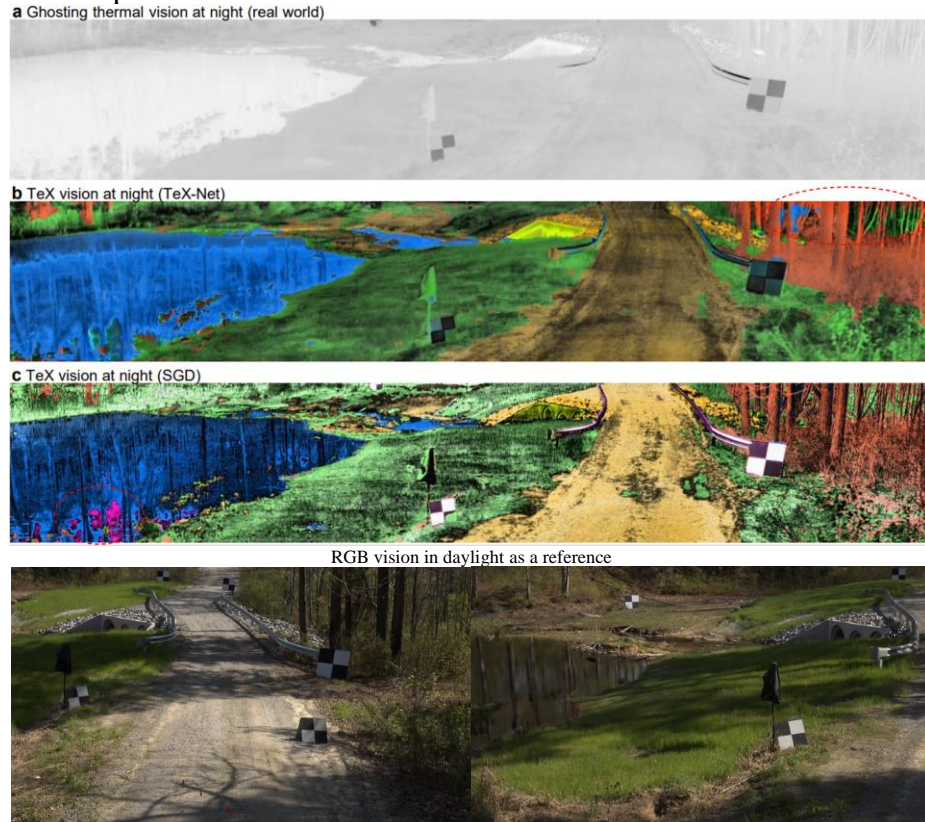


Fig.2 Evidence of wrong material prediction in TeX vision. In Fig.b, trees (brown) are sometimes predicted as water (blue) or vegetation (green). In Fig.c, part of water is predicted as metal (purple), since that part of water is the mirror image of sky and metal in the scene is also reflecting sky signal.

- Segmentation:

In this very first paper of HADAR, our algorithm to convert material map to semantic segmentation (algorithm 4 in the Supple. Info.) is currently a non-machine-learning approach, and HADAR semantics is completely physics driven. Wrong material prediction will immediately hamper HADAR semantics. However, we emphasize that in the future the conversion between material map and semantic segmentation can also be learning based making use of spatial information. Therefore, HADAR semantics would be less influenced by wrong material prediction. As can be seen in the following figure, wrong material prediction leads to <100% mIoU of HADAR semantics, but still HADAR outperforms thermal segmentation.

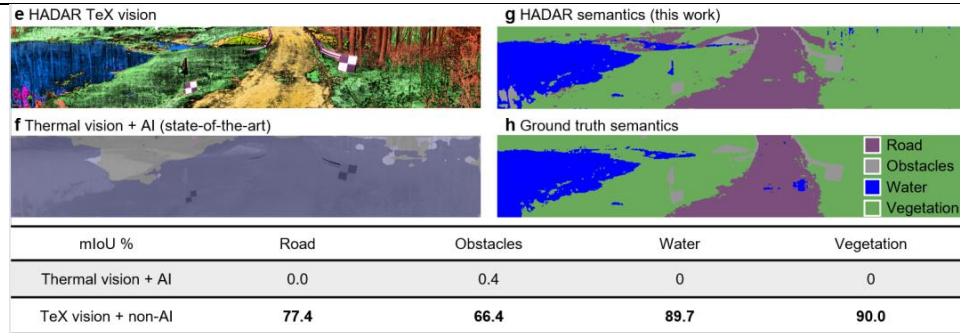


Fig.3 Consequence of wrong material predictions in segmentation

- Detection:**
 In our demonstration of human vs. robot detection, the detection is performed on a certain material region. In the presence of wrong material prediction, the material region will not match the ground truth material region, but detection is robust since detection models will further search for certain spatial patterns in the material region. The detection result shown in Extended Data Fig.7 is cited below as an example. Explicitly, there are a few pixels under the car or around the human leg have wrong material predictions (ground truth: road in purple; prediction: aluminum in green). For robot detection in (d) where image smoothing has been done before detection, the detection model still yields correct result, as the wrong material prediction has not interfered with the spatial pattern of the robot.

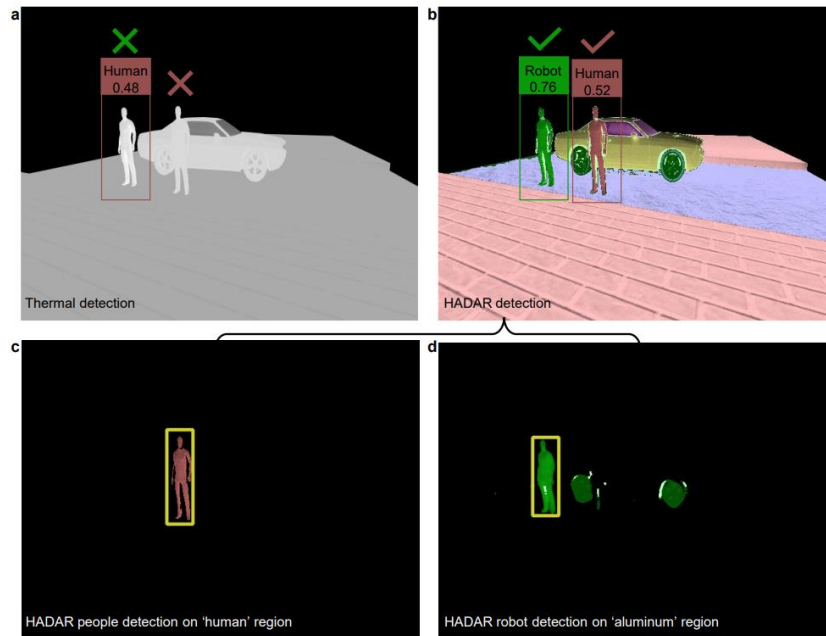


Fig.4 consequence of wrong material predictions in detection

- Ranging:**

We observed that when the predicted TeX vision with wrong material predictions was used for ranging, the ranging accuracy will decrease as compared to that based on the ground truth TeX vision with correct material predictions, as can be seen in table (d) below. But we observed that our argument ‘HADAR ranging at night beats thermal ranging and matches RGB stereovision in daylight’ still holds.

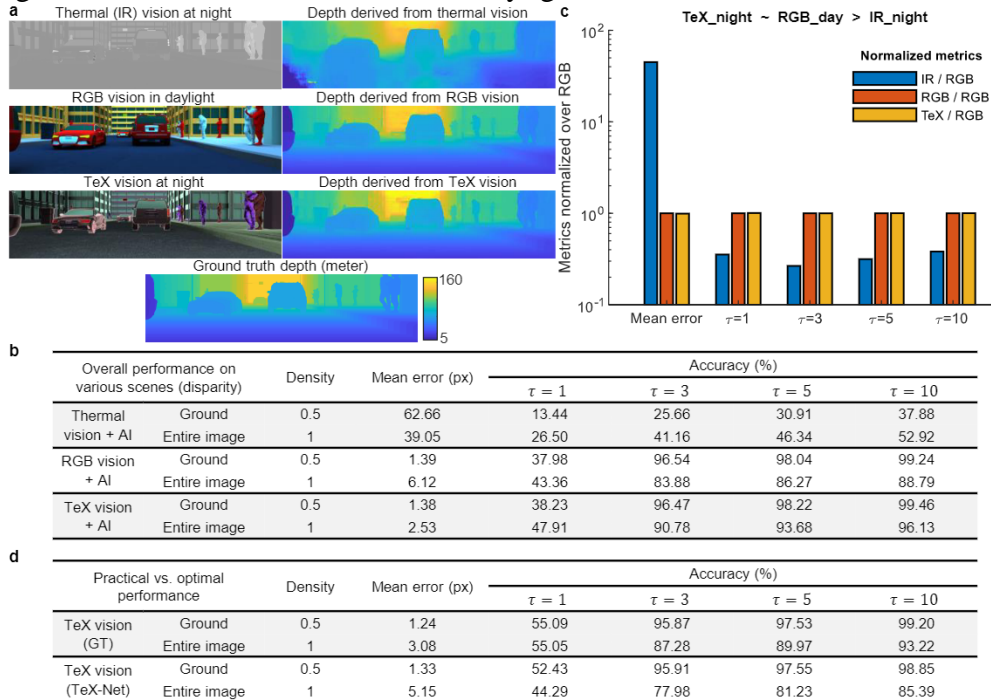


Fig.5 Consequence of wrong material predictions in ranging

Revisions:

Accordingly, we have revised the manuscript to reflect the above points.

<in the caption of newly added Extended Data Fig.3>:

‘...Most of the water pixels can be correctly estimated as ‘water’, except for a small portion corresponding to the sky image which has been estimated as ‘metal’, since metal also reflects the sky signal. TeX-Net utilizes both spatial information and spectral information for TeX decomposition, and hence its TeX vision is spatially smoother. In contrast, TeX-SGD mainly makes use of spectral information and decomposes TeX attributes pixel per pixel. Compared with TeX-Net, we observed that TeX-SGD is better at material identification and texture recovery for fine structures...’

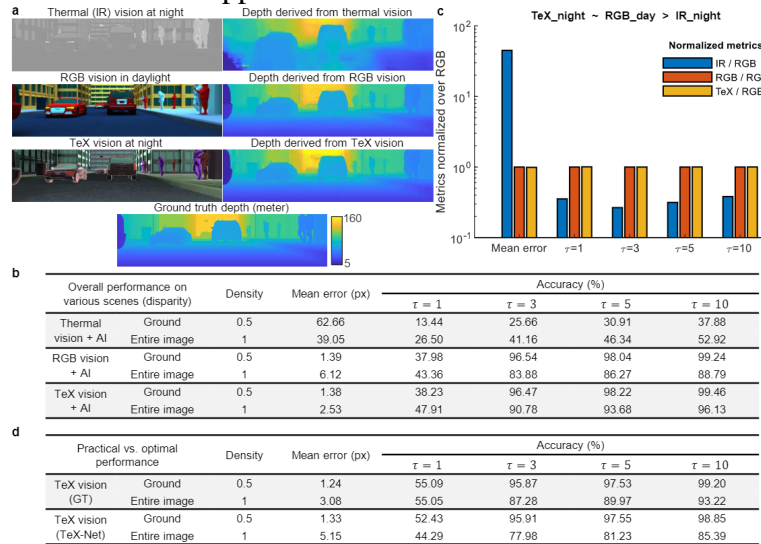
<in the caption of revised Extended Data Fig.8>:

‘...In the future, learning-based approaches to convert material map to semantic segmentation with the help of spatial information can further improve HADAR semantics...’

<in the caption of revised Extended Data Fig.7>:

‘...We also observed that HADAR detection is robust against wrong material predictions, see the few road pixels under the car and around the human leg that have been predicted as ‘aluminum’ in (b) and (d).’

<new Fig.S19, Sec.SIIIA of the Supple. Info.>:



C5 4. In the comparison of this paper, the authors' distinction between humans and robots does not look extensive enough and there should be more visual examples to support the effectiveness of the proposed approach.

R5 We agree with the reviewer that the human-robot identification problem itself is not extensive in the current world we live, but it is expected to be much more extensive by 2030 when human-robot interaction becomes intense. Also, we positively confirm that HADAR does have more effective applications in computer vision even in the current world. Before giving our revisions, we'd like to briefly explain our motivation in selecting the human-robot identification as the pertinent and concise example in the main text.

The advantage of HADAR vs. AI-enhanced thermal imaging comes from the spectral resolution and the disentanglement of temperature (T), emissivity (e), and texture (X). For detection and segmentation, thermal sensing based on radiance (or intensity) is vulnerable to errors as it only exploits the spatial information. This approach suffers from fundamental challenges whenever the spatial information is poor. In the following, we list some common and extensive situations where the spatial information is poor and thermal detection and segmentation is elusive.

- Most off-road scenes where spatial patterns are irregular:
For off-road scenes, there is no regular spatial patterns to learn, and detection/segmentation can only be based on intensity contrast. As demonstrated below as well as in Extended Data Fig.8, segmentation of real-world thermal images based on either AI algorithms (DANet) or non-AI K-means clustering cannot distinguish the road and grass clearly. Using material identification, HADAR achieves off-road semantic segmentation.

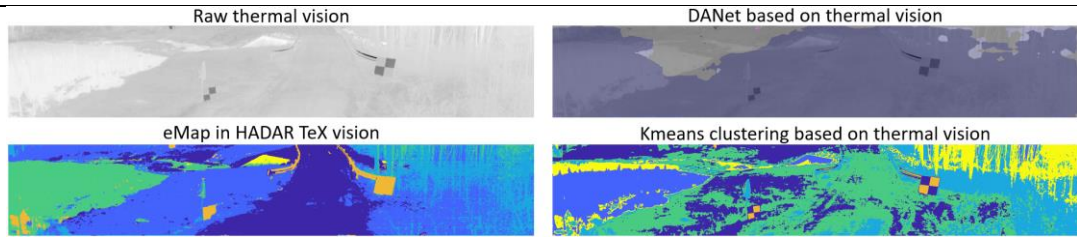


Fig6. HADAR vs. thermal sensing in distinguishing road and grass. HADAR eMap is the map of material index in the HADAR material library. This figure is generated particularly for the reviewer.

All path planning/segmentation based on thermal vision consistently has poor performance, as can be seen in the following test on the Desert scene in our HADAR database.

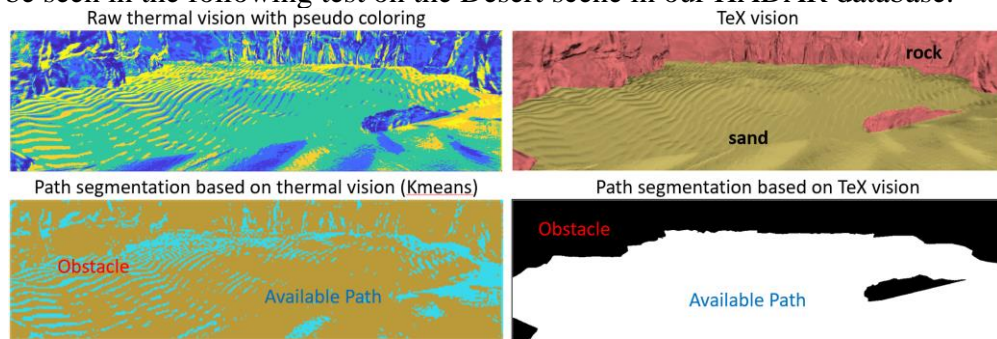


Fig7. HADAR vs. thermal sensing in distinguishing road and obstacles. This figure is generated particularly for the reviewer.

2. Most thermal images are of low-contrast and vague spatial boundaries:

Even for regular spatial patterns, thermal vision with low contrast (exhibiting the ghosting effect) will pose challenges to detection. This is common in almost all thermal images, and it affects the detection, depending on the specific contrast and sharpness of the thermal image. To clearly show this to readers, we have designed scenes with particular parameters to emphasize the detection challenge in thermal vision. As can be seen in our Extended Data Fig.7 (cited below), human body visually merges with the car, since they have similar radiance, and hence the human body was not detected.

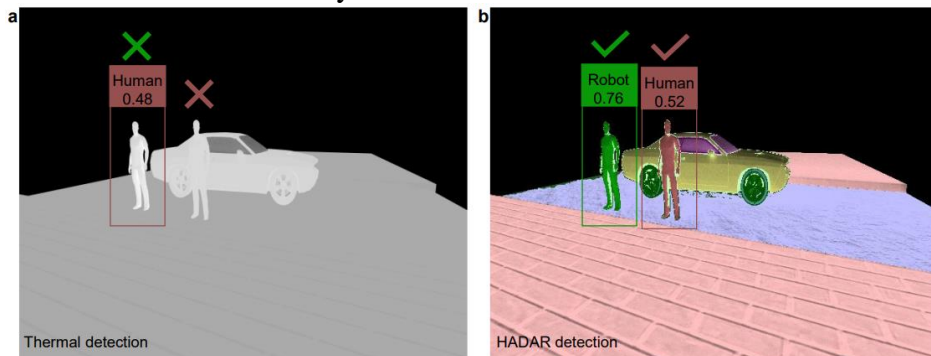


Fig.8 HADAR vs. thermal sensing in detecting low-contrast objects.

3. Visual ambiguity:

	<p>Even with clear spatial patterns and boundaries, visual detection algorithms do not take advantage of underlying material properties or physics-driven context such as temperature and spectral emissivity. This can cause wrong detection. One well-known example is the Phantom braking mentioned in our manuscript. Our human vs. robot identification is another example.</p> <p>In summary, the key argument we want to convey is that the s.o.t.a. segmentation based on spatial patterns alone can result in phenomena such as phantom braking. Other methods using solely total intensity of the signal (e.g., K-means clustering) are inaccurate. In such cases, HADAR exploiting spectral material fingerprint and physics-driven context (e.g., temperature) can substantially improve the performance. To make the physics clear to a broad audience, we decide to discuss concise and typical examples in the main text and give general cases in Extended Data. Therefore, we used the pertinent human-robot identification problem to help illustrate our argument, where we have made the shape and radiance of humans or robots almost the same. We have also provided real-world HADAR performance in extensive off-road scenes in Extended Data.</p> <p>Revisions: In addition to the above new results, we have also revised the manuscript to indicate that the advantage of using material fingerprint in detection/segmentation is extensive. < in the caption of revised Extended Data Fig.8>: <i>‘...This real-world off-road scene is a general example to show the importance of material fingerprint in detection/segmentation. ...’</i></p>
C6	<p>The experiments need improvement. In the paper, the algorithm of Thermal sensing + AI comparison is not state-of-the-art, and the authors should add the comparison with the latest algorithm.</p>
R6	<p>We thank the reviewer for suggesting using the latest algorithms. In this revised manuscript, we have made our best efforts to update the used algorithms to the state-of-the-art.</p> <p>1. <u>Detection:</u> For human and robot detection, we have used <i>the latest algorithm of thermal YOLO</i> (YOLO-v5 fine-tuned on the thermal automotive dataset, https://github.com/MAli-Farooq/Thermal-YOLO-And-Model-Optimization-Using-TensorFlowLite, 2022). This model was chosen because other reviewer(s) also wanted us to use thermal people detectors. The detection result is shown in Extended Data Fig.7, as cited below. We emphasize that the results are consistent as before.</p>

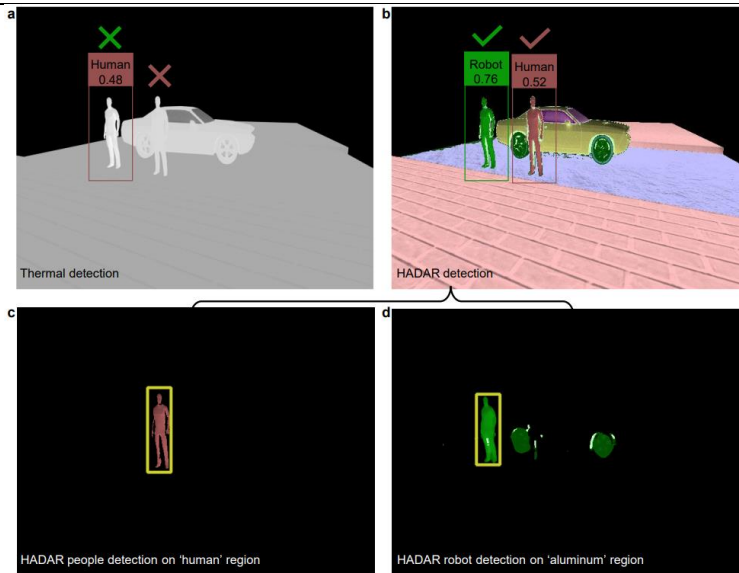


Fig.9 HADAR detection vs. thermal sensing with the latest AI algorithm.

2. Segmentation:

DeepLab is the defacto standard for semantic segmentation (e.g., DeepLabv3 from google [1], 2017). In our manuscript, we used DANet [2] (2019) as the segmentation model which is more recent than DeepLab. Another main reason we used DANet is that we are utilizing both spatial and spectral information in our TeX-Net to generate HADAR TeX vision. It is therefore pertinent to also use a Dual-Attention Network for thermal sensing for comparison.

3. Ranging:

- For monocular depth estimation, we used *the latest algorithm GCNDepth* [3] (2023). Another main reason we chose GCNDepth is that it is a convolutional model and can be easily applied to our HADAR database which has different image sizes from their training set.
- For binocular stereovision, we used the sub-pixel block matching algorithm (non-machine-learning) to demonstrate our fundamental limit of ranging, and we used DeepPruner [4] (2019) as the AI algorithm for HADAR ranging statistics. Similar to sub-pixel block matching, DeepPruner was chosen because it relies on differentiable PatchMatch, which has close relation to our theory of texture recovery and stereo matching. In the very first paper of HADAR, we think it's beneficial to choose models that are close or analogous to our theory so as to make the physics clear.

We agree with the reviewer that in the future all algorithms should be updated constantly to the latest to test HADAR performance vs. thermal sensing.

Reference(s):

[1] Florian, L-CCGP, and Schroff Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation." Conference on computer vision and pattern recognition (CVPR). IEEE/CVF. Vol. 6. 2017.

	<p>[2] Fu, Jun, et al. "Dual attention network for scene segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.</p> <p>[3] Masoumian, Armin, et al. "Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network." Neurocomputing 517 (2023): 81-92.</p> <p>[4] Duggal, Shivam, et al. "Deepruner: Learning efficient stereo matching via differentiable patchmatch." Proceedings of the IEEE/CVF international conference on computer vision. 2019.</p>
C7	<p>6. Some experimental results seem unconvincing. For example, in Extended Data Fig. 8, why choose 10 frames of the left camera in the city block dataset 1? Can you evaluate them on entire dataset?</p>
R7	<p>We regret that our statistics was not explained well in the previous version of the manuscript. Previously, the segmentation was done on the CityBlock dataset (called Street-Long-Animation now). The dataset has left and right views, and each view has 100 frames. The dataset was split into training set (90 frames/view) + validation set (10 frames of the right view, 8:10:98) + test set (10 frames of the left view, 8:10:98). Validation and test frames were so chosen that they can expand to the whole dataset and would have the same diversity as the training set. The performance was evaluated and presented on the test set, which is a standard way of doing statistics. We should have said that the statistics was done on test set, instead of saying on 10 frames of the left camera.</p> <p>Nevertheless, other reviewer(s) have pointed out that it is better to include cross validation. <u>Therefore, in this revised manuscript, we have used 5-fold cross validation for both ranging and segmentation statistics.</u> TeX vision of all frames, generated when they are in the validation set, was used in statistics. We believe this is a more convincing answer to this comment as well.</p>
C8	<p>7. The fusion of infrared image and visible image perception is a common practice in real scenes. The authors should compare the performance after fusing optical imaging with raw thermal vision and HADAR TeX vision.</p>
R8	<p>We agree with the reviewer that the fusion of thermal images and optical images has been a common practice in real scenes. In the literature [1-4], thermal images have been fused with optical images possessing poor ambient illumination for night-vision enhancement. The goal is to integrate complementary information from different sensors, i.e., the detailed textures in optical images and target highlighting in thermal images.</p> <p>Here, we would like to first compare HADAR vs. thermal-infrared fusion and clarify our objective underlying this work before we revise to incorporate the reviewer's comment.</p> <ol style="list-style-type: none"> <u>HADAR is fully-passive.</u> Our focus is to work in the extremely low light regime where conventional RGB sensors record no information at all. The multi-sensor fusion approach may have comprehensive information as HADAR does but cannot be fully passive. Explicitly, the visible-infrared image fusion is generically pseudo-passive as visible images rely on at least some level of ambient illumination. For completely dark scenes, such as our real-world off-road scene at night, there is absolutely no information in visible images and hence the visible-infrared image fusion is not better than the raw thermal vision. <u>Our work aims to demonstrate that we can get rich information solely out of the heat signal which was previously thought to be impossible.</u>

In our work, our focus is to demonstrate the advantage of HADAR vs. thermal sensing which are both based on heat signal. When full passivity or scalability is not required and more sensors can be considered in a multi-sensor fusion approach, HADAR can replace the traditional infrared sensor and work together with other sensors like the visible RGB camera. Explicitly, to combine RGB with TeX, one can following the procedures below. (1) Convert RGB images to grayscale. That is, keep the textures from material response in the visible-light range (which is complementary to textures in the infrared range), and discard the color. The color from TeX vision will be adopted since that has physics-driven semantic meanings. (2) Fuse X channel with grayscale optical images. (3) Use the fused image to replace the original X and, together with T and e channels, form the new ‘enhanced’ TeX vision images. The following figure demonstrates the RGB-TeX fusion in comparison with RGB-IR fusion. Image fusion is implemented with the benchmark algorithms [5]. Since TeX vision is better than thermal IR vision, it is also seen that the new ‘enhanced’ TeX vision (TeX + RGB fusion) is better than visible-infrared image fusion (IR + RGB). For example, the textures on the street and pavement in TeX-RGB fusion are more than that in IR-RGB fusion.



Fig.10 TeX+RGB fusion vs. IR+RGB fusion

3. HADAR is physics aware.

	<p>HADAR has the material/semantic information in the e channel that visible-infrared image fusion does not have. Also, HADAR temperature in the T channel is more accurate than traditional thermography.</p> <p>4. <u>The main goal of HADAR is to pursue night-time stereovision with performance metrics as good as the RGB stereovision in daylight.</u></p> <p>Visible-infrared image fusion at night is to enhance the RGB vision with thermal vision for the part of image with poor ambient illumination. To better support our argument, instead of comparing with visible-infrared image fusion, we directly compare our TeX vision at night (pitch-darkness) with RGB vision in daylight and show that HADAR has comparable textures as well as ranging accuracy. We believe this is an even stronger comparison.</p> <p>Revisions: <i>Accordingly, we have added the above analysis and the above figure in Sec. SVE and Fig.S34 of the Supple. Info.</i></p> <p>Reference(s): [1] Gu, Yansong, et al. "Advanced driving assistance based on the fusion of infrared and visible images." Entropy 23.2 (2021): 239. [2] Li, Hui, Xiao-Jun Wu, and Josef Kittler. "RFN-Nest: An end-to-end residual fusion network for infrared and visible images." Information Fusion 73 (2021): 72-86. [3] Zhao, Zixiang, et al. "Bayesian fusion for infrared and visible images." Signal Processing 177 (2020): 107734. [4] Zhou, Zhiqiang, et al. "Fusion of infrared and visible images for night-vision context enhancement." Applied optics 55.23 (2016): 6480-6490. [5] X. Zhang, P. Ye, G. Xiao. VIFB: A Visible and Infrared Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.</p>
C9	<p>8.In open environments, there exist unknown classes of materials and unknown scenes. I wonder how the TeX vision would work in open environments.</p>
R9	<p>We thank the reviewer for mentioning the HADAR performance in open environments with unknown materials and scenes. Other reviewer(s) also have comments regarding real-world experiments. Therefore, in this revised manuscript, we have made our best efforts to provide the HADAR prototype-2 experiments in real-world off-road scenes at night, see Fig.6 and Extended Data Figs. 3-8 for results.</p> <p>As shown below, the off-road scene is an open environment with complexity and diversity in details. Except for a few man-made objects (e.g., the checkerboard marks and the culverts) and a few that we can visually and roughly tell (e.g., water, grass, trees), there are many unknown details in the scene. The experiment was conducted under the DARPA IH project (The Defense Advanced Research Projects Agency, Invisible Headlights). The material library was not collected, and we don't exactly know how many materials there are in the scene. The sky (which is a significant environmental object and important to our HADAR theory) was not directly observed in the image as well. The sensor was a pushbroom hyperspectral imager that has horizontal streak noise (sensor noise has been detailed in Methods—prototype HADAR calibration). We believe this is a very general experiment to prove real-world HADAR efficacy. In the following, we briefly describe how we obtained the TeX vision and HADAR detection and ranging. More details have been given in Sec. SV of the Supple. Info.</p>

1. TeX vision:

We first used a customized TES (temperature-emissivity separation) algorithm followed by K-means clustering to estimate the material library. To be distinguished from the exact ground truth material library, here we call the estimated material library the semantic library. The reason is also related to the fact that there are considerable unknown details in the scene. For example, in this real-world off-road scene, the gravel road by a grass lawn may consist of soil, sand, or little stones that cannot be spatially resolved by sensor pixels. In this case, each road pixel may exhibit a slightly different spectral emissivity curve and the ground truth is unknown, but their emissivity curves are still distinct with the grass, trees, or water. Therefore, using averaged emissivity curves can capture the semantics of road vs. others, while the deviation from the exact emissivity will become a perturbation to the temperature and thermal lighting factors, or remain in the physics-based loss, “res”. This error will diminish as the number of semantic categories in the library increases. We emphasize that a semantic library by no means covers all exact material in the scene. K-means clustering needs an input of number of clusters, and that determines the dimension of the semantic library. We read the signal off the reflecting part of the checkerboards to estimate the sky radiance. With the above approximation, we used our TeX-SGD (semi-global decomposition) to generate TeX vision, and this TeX vision was used to estimate the ground truth TeX vision to train TeX-Net. TeX vision results are shown below.

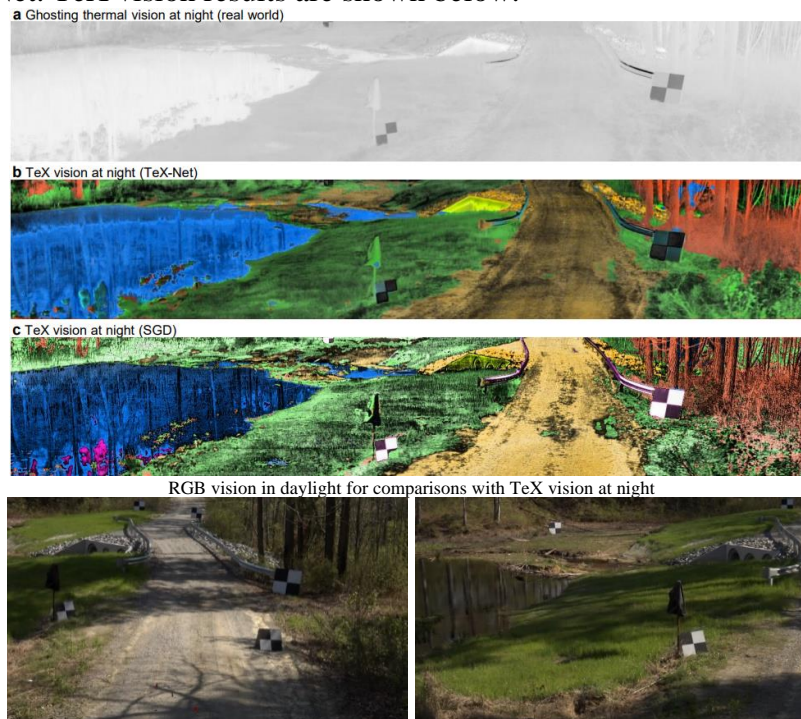


Fig.11 TeX vision vs. IR and RGB vision for open environments with unknown materials

In the above results, TeX-SGD only used the spectral information and physics loss for TeX vision, while TeX-Net utilized both spatial and spectral information. We observed that TeX-SGD is better at material identification and texture recovery for fine structures, such as bark

wrinkles, culverts, bridge fence, etc., and TeX-Net is better at spatial smoothness. Both TeX vision results demonstrate that HADAR can see through the darkness as if it were day.

2. HADAR detection (segmentation):

With the above TeX vision, we used a custom algorithm to convert the material map to semantic segmentation. As discussed in Reply R4, there are few pixels where material prediction is wrong, but the mIoU statistics confirms that HADAR has captured the semantics reasonably well.

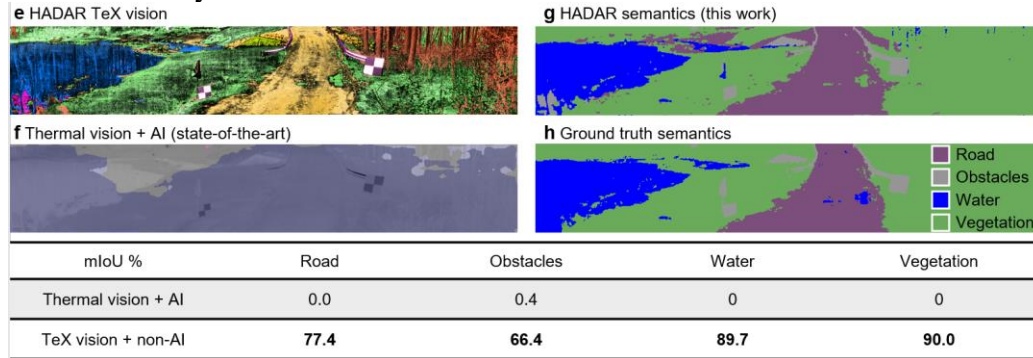


Fig.12 HADAR segmentation for open environments with unknown materials

3. HADAR ranging:

With the above TeX vision, we used a monocular depth estimation model, GCNDepth, for HADAR ranging. The ranging statistics (Fig.f below) clearly shows that HADAR ranging at night beats thermal (IR) ranging and is comparable to RGB stereovision in daylight. This demonstrates our argument ‘HADAR sees texture and depth through the darkness as if it were day’.

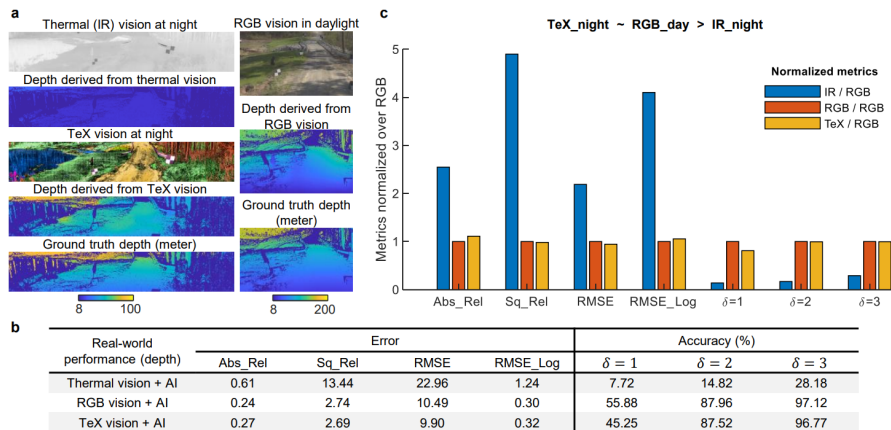


Fig.13 HADAR ranging for open environments with unknown materials.

We thank the reviewer once again for the time and efforts spent to provide us such details comments. With the above changes, we believe the manuscript is now significantly improved and ready for publication.

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The reviewer commends the diligent and meticulous approach taken by the authors. Their level of effort in addressing the reviewer's comments is commendable. The supplementary material comprehensively covers the mathematical foundations of formulating the fully-passive and physically aware machine perception. The paper has rich details on the physics and mathematical modeling of the imaging process. Given that the author's intended takeaway from this work is the breakthrough performance of computer vision tasks with Tex Vision compared to thermal, the paper needs to be fundamentally strong in ML components and provide rich details in these sections. To this point, the reviewer suggests improvements in certain important details of ML that are necessary to fully validate the authors' claims:

1. Please thoroughly check the Error values, particularly Sq_Rel, RMSE, and RMSE_Log, for correctness in Fig 6b and c. For example, RMSE for Tex+AI is lower than RGB+AI; however, RMSE_Log for RGB+AI is lower than Tex+AI. Per the reviewer's understanding, all the error entities for RGB+AI should be lower than TEX+AI. Also, What do the delta parameters presented in the same table correspond to in the caption and discussions?
2. On page 9 of the main body, under the methods section. The authors state, "The dimensionality curse for high spectral resolution (536 bands used) leads machine learning to over- fitting, and slight deviation between Monte Carlo simulation and theoretical prediction can be observed in Fig. 3b. Once the dimensionality curse is relieved, perfect agreement can be reached." How precisely do the authors relieve this dimensionality curse?
3. The authors' original supposition was developing low-cost hardware with the HADAR paradigm. While the efficacy looks promising on synthetic examples and very high-cost sensors, the results call into question the efficacy of the approach utilizing the original low-cost systems upon which the original supposition was based and the feasibility of integration in real-world applications. In keeping with the original premise, it would be helpful to readers to present a visual comparison of the TEX vision side-by-side, achieved via prototypes 1 and 2, to understand the differences in HADAR's performance at different cost settings.
4. The authors have done an excellent job discussing underlying physics and generating data simulations based on these physics models in the supplementary materials section. However, the ML/DL sections need to be strengthened with associated details as they only have minimal information in the supplementary material.
5. On page 41 of the supplementary material, the gain in texture density with Tex-Vision compared to thermal vision is seemingly quite small. Also, is there a way to quantify if the texture metric is also impacted by noise? For example, if the TEX vision has significant noise in texture compared to

thermal, that would also result in higher texture density. It would also be useful to readers if information about the common value for standard RGB images is provided.

6. It would be useful to the readers if the details on datasets utilized for benchmarking detection, ranging, and segmentation with Texnet were provided. Also, the number of images used for training, testing, and validation for the experiments need to be included as well as the training strategy used.

7. In figure s19, what does tau correspond to?

8. In fig s24, the claim about human vs. robot identification compared to thermal, as seen from the supplementary material, is due to the advantages provided by multi-spectra. As shown in Fig S24, the ability to perform semantic segmentation exploiting the material signature to mask the underlying different material layers followed by detection in that masked layer raises concern about the simultaneous detection of different objects corresponding to different spectra in a scene and performing end-to-end semantic segmentation. Given this strategy, one can directly utilize the multispectral tensor data for this task without TeX-vision, where each spectra can be utilized separately to identify the subject composed of given material (skin vs. aluminum). To address this concern, can TEX vision enable the simultaneous detection of multi-objects from a single frame without masking?

9. On page 71 of supplementary materials, estimating the signatures on the fly for the categorization of materials sounds like a good idea. As the Authors utilize k means to estimate the categories, how do they choose the k as this impacts the categorization process leading to over/underfitting noise during categorization? An incorrect categorization of materials can lead to errors in semantic representations, which propagate to successive tasks. Also, what is a customized TES algorithm?

10. Since the authors utilized the result of Tex-SGD as ground truth labels to train the Tex-net, the DL trained is fundamentally limited by the performance of Tex-SGD. Consequently, it is not a fair comparison of the performance of TexNET and TexSGD when the performance of one constrains the performance of the other. The authors should develop alternate loss functions to decouple the dependency or augment some additional form of label for training the Texnet in a supervised fashion.

Referee #3 (Remarks to the Author):

The authors have conducted extensive experiments to demonstrate the effectiveness and efficiency of the proposed method. However, I have some minor comments and suggestions that I hope the authors can address before publication. Following are a few suggestions and some questions for the authors:

1. Most of the scenes in the demo examples provided are relatively easy to segment. I am very curious about the performance in real-world scenes where objects are overlapped or occluded.
2. The authors provide a video about the real-world experiment. But the synthesis results are not stable as seen in Tex.avi, for example the trees in the upper right corner of the video keep flashing.
3. The video of the real-world experiment provided by the authors is short in length and lacks more challenges such as the movement of the objects. Authors are encouraged to provide more video visualizations containing more diversity of scenes.
4. Considering that the authors use the ability of object detection to evaluate the quality of TeX vision, the authors should further introduce evaluation of visual object tracking to demonstrate the robustness of TeX vision.

Author Rebuttals to Second Revision:

Reviewer 1	
C0	The reviewer commends the diligent and meticulous approach taken by the authors. Their level of effort in addressing the reviewer's comments is commendable. The supplementary material comprehensively covers the mathematical foundations of formulating the fully-passive and physically aware machine perception. The paper has rich details on the physics and mathematical modeling of the imaging process. Given that the author's intended takeaway from this work is the breakthrough performance of computer vision tasks with Tex Vision compared to thermal, the paper needs to be fundamentally strong in ML components and provide rich details in these sections. To this point, the reviewer suggests improvements in certain important details of ML that are necessary to fully validate the authors' claims:
R0	We thank the reviewer for all the efforts and patience throughout the whole review process. Your encouraging response, detailed comments and insights have helped us improve the work and converge to this final version for publication. We have addressed each comment individually below and made corresponding revisions to improve the quality of this manuscript.
C1	Please thoroughly check the Error values, particularly Sq_Rel, RMSE, and RMSE_Log, for correctness in Fig 6b and c. For example, RMSE for Tex+AI is lower than RGB+AI; however, RMSE_Log for RGB+AI is lower than Tex+AI. Per the reviewer's understanding, all the error entities for RGB+AI should be lower than TEX+AI. Also, What do the delta parameters presented in the same table correspond to in the caption and discussions?
R1	<p>We would like to address this comment in two aspects.</p> <ol style="list-style-type: none"> 1. The reviewer touched upon an interesting question: is RGB+AI during the day always better than TeX+AI at night? Our results show that the answer to this question is negative. Indeed, we used RGB vision as a reference in this work, and we explained that HADAR recovers textures by reconstructing the scattering heat signal, mimicking daylight imaging. That means the mechanism of imaging textures in RGB and TeX is the same, i.e., retrieving the scattering signal. However, a materials' response to light (spectral features of emissivity/reflectivity) in the visible-light spectrum is different from that in the thermal infrared spectrum. In addition, the wavelength of the thermal infrared is around one order in magnitude larger than that of the visible light. These factors will result in slightly different textures in thermal and optical spectral ranges. Moreover, RGB vision only has 3 channels, but HADAR prototype-2 in this work has 256 channels. All these factors can lead to more textures in TeX images than RGB images, depending on specific scenes. Therefore, we only claimed a conservative 'comparable' ranging performance between RGB+AI and TeX+AI, as we have observed, instead of an absolute performance ranking. <p><u>We have added the above analyses in Sec.SID and Sec.SVD of the Supple. Info., as cited below.</u></p> <p style="text-align: center; font-size: small;">However, we remind that materials' response to light (spectral features of emissivity) in the visible-light spectrum is different with that in the thermal infrared spectrum. Furthermore, the wavelength of the thermal infrared is around one order in magnitude larger than that of the visible light. They also result in different textures in thermal images and optical images.</p>

D. Texture comparison and analysis between TeX vision and RGB vision in experiments

As explained in Sec. SID, textures in RGB vision and TeX vision images will be different, due to different working wavelength and different material responses in these two spectral

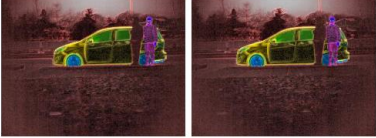
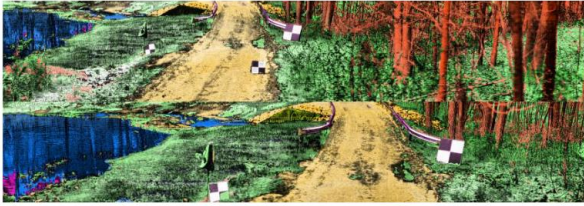
73

ranges. Furthermore, the following factors will also lead to different textures in two different imaging modalities. (1) Related to the working wavelength, the pixel size of a thermal infrared sensor is of the order of $20\mu\text{m}$, while the pixel size of an RGB camera is below $2\mu\text{m}$. The $\sim 10\times$ difference in working wavelength and pixel size leads to a poorer spatial resolution and less fine textures in TeX vision. (2) The electronic noise (NEP) of state-of-the-art thermal sensors is much higher than that of state-of-the-art RGB cameras. This means RGB images usually have a higher signal-to-noise ratio and more subtle textures. Especially in hyperspectral imaging, there is systematic noise like horizontal streaks in images for ‘pushbroom’ sensors, which will pollute real textures. (3) The state-of-the-art hyperspectral imagers are much slower than regular RGB cameras. The former takes several seconds to form one image, while the latter takes only milliseconds. Motion blur in real-world scenes becomes severer in current TeX vision than RGB vision. (4) The state-of-the-art hyperspectral imagers are usually focal-plane arrays, which means the sensor is focusing at infinity. While RGB cameras can focus on the surrounding scenes, focus blur becomes severer in current TeX vision than RGB vision. All these factors have been observed in the TeX vision obtained in real-world experiments, as shown in Fig. S33.

2. We thank the reviewer for the reminder to check metric values. We have double checked them and we confirm that the statistical results are correct. In fact, after checking relevant literature in computer vision, we realize it is a common phenomenon that a certain algorithm favors one metric while another algorithm may favor others. This usually occurs especially when two competing approaches (Day- RGB + AI, Night -TeX vision + AI) have comparable performances. Explicitly, it is possible that different depth metrics, due to different mathematical definitions, show different relative rankings, see, for example, Tab. V and Tab. VIII of Ref. [1]. In turn, this indicates that the ranging performance between Daytime:RGB+AI and Nighttime:TeX+AI is comparable, consistent with our argument.

In this revised version, we have added a section in Methods to clearly define all metrics, as cited below.

	<p>Standard depth metrics Let $pred$ and gt denote predicted and ground truth depth, respectively. D represents the set of all predicted depth values. \cdot returns the number of elements, and $\ \cdot\$ returns the absolute value. The standard depth metrics used in Fig. 6 are defined as below.</p> <p>Absolute and Relative Error,</p> $Abs_Rel = 1/ D \cdot \sum_{pred \in D} \ gt - pred\ /gt.$ <p>Squared Relative Error,</p> $Sq_Rel = 1/ D \cdot \sum_{pred \in D} \ gt - pred\ ^2/gt.$ <p>Root Mean Squared Error,</p> $RMSE = \sqrt{1/ D \cdot \sum_{pred \in D} \ gt - pred\ ^2}.$ <p>Root Mean Squared Log Error,</p> $RMSE_Log = \sqrt{\frac{1}{ D } \sum_{pred \in D} \ \log(gt) - \log(pred)\ ^2}.$ <p>δ_t Accuracy,</p> $\delta_t = \frac{1}{ D } \{pred \in D \max(\frac{gt}{pred}, \frac{pred}{gt}) < 1.25^t\} .$ <p>Reference(s): [1] Masoumian, Armin, et al. "GCNDepth: Self-supervised monocular depth estimation based on graph convolutional network." <i>Neurocomputing</i> 517 (2023): 81-92.</p>
C2	<p>On page 9 of the main body, under the methods section. The authors state, "The dimensionality curse for high spectral resolution (536 bands used) leads machine learning to over- fitting, and slight deviation between Monte Carlo simulation and theoretical prediction can be observed in Fig. 3b. Once the dimensionality curse is relieved, perfect agreement can be reached." How precisely do the authors relieve this dimensionality curse?</p>
R2	<p>We regret that the statement was not made clear in the previous version. In this revised version, we have explicitly added that we use only 3 spectral bands for both theory and machine learning for human-robot identification to relieve the dimensionality curse, as cited below.</p> <p style="text-align: center;">Fig. 3b. Once the dimensionality curse is relieved, perfect agreement can be reached, see Fig. S7 in Supple. Info. where all spectra are down-sampled into 3 spectral bands (dimension = 3) for both theory and machine learning.</p>
C3	<p>The authors' original supposition was developing low-cost hardware with the HADAR paradigm. While the efficacy looks promising on synthetic examples and very high-cost sensors, the results call into question the efficacy of the approach utilizing the original low-cost systems upon which the original supposition was based and the feasibility of integration in real-world applications. In keeping with the original premise, it would be helpful to readers to present a visual comparison of the TeX vision side-by-side, achieved via prototypes 1 and 2, to understand the differences in HADAR's performance at different cost settings.</p>
R3	<p>We thank the reviewer for the helpful suggestion. <u>In this revised version, we have added a new section (Sec.SVF) and a new figure (Fig.S35) in the Supple. Info. about the visual comparison of the TeX vision obtained by prototypes 1 and 2, as cited below.</u></p>

	<p>F. TeX vision comparison between two HADAR prototypes</p> <p>Here, Fig. S35 shows the visual comparison of TeX vision obtained by our two HADAR prototypes for night scenes. This provides the intuitive understanding of TeX vision with different sensor performance and cost settings.</p> <p>TeX vision by HADAR prototype-1</p>  <p>TeX vision by HADAR prototype-2</p> 
C4	<p>The authors have done an excellent job discussing underlying physics and generating data simulations based on these physics models in the supplementary materials section. However, the ML/DL sections need to be strengthened with associated details as they only have minimal information in the supplementary material.</p>
R4	<p>We regret that the details about ML/DL were distributed in multiple places and not presented systematically for readers' convenience. <u>In this revised version, we have expanded Sec.SIIIA of the Supple. Info. to fully explain our TeX-Net and machine learning, as cited below. Details about the HADAR database (training data), number of images used for training and validation (5-fold cross validation), hyper-parameters like learning rate etc. (training strategy) are all included, in addition to the Saliency maps and ML/DL performance. We have also provided detailed README for both HADAR and TeX-Net on the GitHub pages (will be public after publication), as cited below.</u></p> <p style="text-align: center;">SIII. HADAR estimation theory II: inverse mapping in applications</p> <ul style="list-style-type: none"> A. TeX-Net and machine learning <ul style="list-style-type: none"> 1. Training data and training strategy 2. Saliency maps 3. Performance and training loss B. Analytical inverse functions, Least-squares estimator, and the TeX-SGD (Semi-Global Decomposition) C. AGC on TeX vision D. Pseudo-TeX vision E. Physics-driven semantic segmentation, object detection and visual object tracking

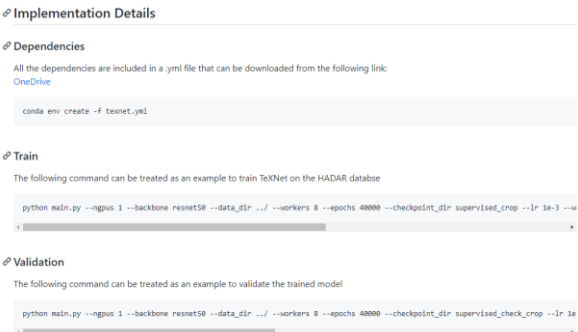
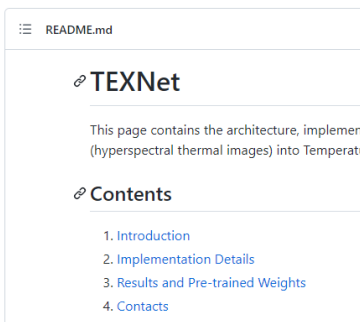
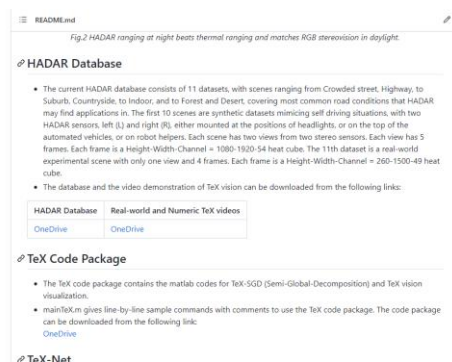
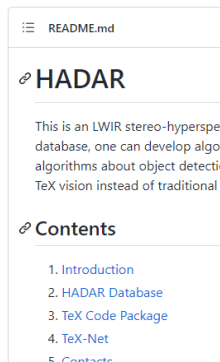
1. Training data and training strategy

Our TeX-Net was trained on the HADAR database (<https://github.com/FanglinBao/HADAR>). The HADAR database includes dissimilar scenes like Crowded Street, Highway, Suburb, Countryside, Indoor, Forest, Desert, etc., covering most common road conditions that HADAR may find applications in. The 11th dataset is a real-world off-road scene with heat cube dimension Height \times Width \times Channel = 260 \times 1500 \times 49, while the first 10 scenes are synthetic with heat cube dimension Height \times Width \times Channel = 1080 \times 1920 \times 54. The channels in the real-world scene correspond to the 5th ~ 53th channels of the synthetic scenes. The HADAR database mimics self-driving situations, with the HADAR sensor(s) either mounted at the positions of headlights, or on the top of the automated vehicles, or on robot helpers. Each scene has 5 frames for each camera, and there are 30 different kinds of materials in total in the HADAR database. For the Street, Suburb, Rocky Terrain, and the Real-World Off-Road scenes, TeX, RGB and IR images are provided for the purpose of ranging. The Street scene has a long animation version (100 frames, 12 channels). For the real-world experimental scene, HADAR sensor is a pushbroom hyperspectral imager that can produce 256 spectral bands. The heat cubes have been interpolated into 49 channels to match the channels in synthetic scenes. Only 49 channels of all the scenes are used

to train TeX-Net. Full technical details about the HADAR database, such as, ray depth, field of view, material properties, and so on, are available in the readme file along with the database.

We split the HADAR database (11 scenes) into training set (80% data) + validation set (20% data) to train the TeX-Net with 5-fold cross validation. Due to limited experimental data, we manually split the database, instead of randomly splitting, to ensure the same diversity of the validation set and training set. Explicitly, in each fold, one frame per view of each scene was selected for validation. We used a hybrid loss with half supervised loss and half physics loss, and we trained TeX-Net for 40K epochs. Since the real-world scene (260*1500) has a different image size with the synthetic scenes (1080*1920), we used random crop (256*256) in training. The network was trained using the number of workers of 8 and a batch size of 20. The learning rate started at 0.001 and dropped by a factor of 10 at 30000 and 37000 epochs. ADAM optimizer was used with the default momentum parameters. The used ResNet50 model was pre-trained on the ImageNet dataset. For synthetic scenes, ground truth temperature and material are synthesized along with the heat cubes. Thermal lighting factors are solved out with least-squares fitting as the ground truth. For the experimental scene, we first applied our proposed TeX-SCD (semi-global decomposition) to generate the TeX vision, as an estimation of the ground truth TeX vision. TeX-SCD results are then used together with synthetic data to train the TeX-Net. TeX-SCD is a non-machine-learning approach that decomposes TeX pixel per pixel based on the physics loss and a smoothness constraint. The hardware environment was Nvidia RTX A6000 48GB GPU. The TeX-Net codes, pre-trained weights and loss curves are available at <https://github.com/FanglinBao/HADAR/tree/main/TeXNet>.

-----below are screenshots of the GitHub page-----



C5

On page 41 of the supplementary material, the gain in texture density with Tex-Vision compared to thermal vision is seemingly quite small. Also, is there a way to quantify if the texture metric is also impacted by noise? For example, if the TEX vision has significant noise in texture compared to thermal, that would also result in higher texture density. It would also be useful to readers if information about the common value for standard RGB images is provided.

R5

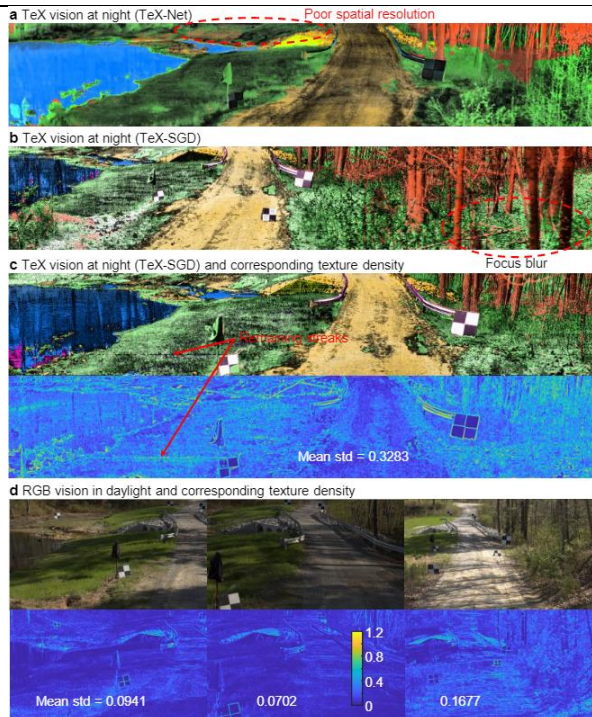
We would like to clarify with some additional details.

1. The absolute gain of texture density with TeX vision compared with the state-of-the-art thermal vision (pseudo coloring) is scene and sensor dependent. This is related to the fact that the absolute ranging accuracy improvement of TeX with respect to IR is scene

dependent. In the specific experiments of Figs.S13 and S14, the mean texture density (standard deviation metric) for enhanced thermal vision is around 0.02, while the mean texture density for TeX vision is around 0.08. The gain is around 4 folds.

2. In this paper, we have introduced two texture metrics, see Sec.SIID of the Supple. Info. (1) The Fisher information metric is immune to noise and is fundamentally related to ranging error, see Fig.S12 of the Supple. Info. However, computing the Fisher information metric requires the knowledge of the ground truth scene and hence is difficult to use in real-world experiments. This is a common phenomenon in estimation theory. (2) The standard deviation (std) metric is easy to use but can be impacted by noise. Note that noise may increase the computed std texture but cannot increase ranging accuracy. Our experimental ranging statistics in Fig.6 of the main text showing 2~5 folds of accuracy improvement are qualitatively consistent with the standard-deviation texture analysis.
3. At last, we thank the reviewer for the useful suggestion about RGB texture quantification. In this revised version, we have improved Fig.S33 of the Supple. Info. to include texture quantification of RGB images, as cited below. Roughly speaking,
 - the mean texture density of thermal vision in our experiments is around 0.02;
 - the mean texture density of TeX vision (HADAR prototype-1) is around 0.08;
 - the mean texture density of RGB vision is around 0.11 on average;
 - the mean texture density of TeX vision (HADAR prototype-2) is around 0.33 for the shown frame in Fig.c and is around 0.31 once averaged over all frames.

It is interesting to see that TeX vision of prototype-2 has a larger texture density than RGB vision. Possible reasons for that include (1) the difference of HADAR sensor noise and the RGB camera noise as explained in the above point 2, (2) HADAR has 256 spectral bands and is much larger than RGB cameras which has only 3 bands, as explained in Reply-R1-point-1, (3) RGB vision suffers from poor ambient illuminations and shadows, and (4) HADAR and the RGB camera have different fields of view. A deeper and thorough analysis deserves extensive future studies. We have provided the above analysis in the context of Fig.S33, as cited below.



It can be seen in Fig. S33c-d that TeX vision can even have a larger texture density than original RGB vision. When the same amount of noise of the HADAR data is introduced into RGB images, and when the RGB images are down-sampled to match HADAR spatial resolution, the mean texture density of the RGB images changes to 0.0975, 0.0624 and 0.1553, respectively. The observation of 'TeX vision has more textures than RGB vision' still holds. Possible reasons for more textures in TeX vision include that (1) remaining HADAR sensor noise in the heat cube gets amplified in generating the TeX vision. (2) Poor ambient illumination (shadow) exists in RGB images. (3) HADAR sensor and RGB cameras have different field of view. And (4) More textures may come from more spectral bands in HADAR than RGB cameras. The last case suggests that it may be possible for HADAR ranging at night to even beat RGB stereovision in daylight. Deeper analysis and verification deserve extensive future studies.

C6	It would be useful to the readers if the details on datasets utilized for benchmarking detection, ranging, and segmentation with Texnet were provided. Also, the number of images used for training, testing, and validation for the experiments need to be included as well as the training strategy used.
R6	We thank the reviewer for this suggestion. As explained in Reply R4, we have now provided the details about the database and machine learning in Sec.SIII of the Supple. Info.
C7	In figure s19, what does tau correspond to?
R7	We apologize for the missing definition. <u>In this revised version, we have added in the caption of Fig.S19 the definitions of all used metrics including tau, as cited below.</u>

	<p>FIG. S19. General HADAR ranging performance over various scenes. (c) corresponds to the ground in table (b). The metrics of TeX comparable with RGB and beating IR demonstrates that HADAR ranging at night beats thermal ranging and is comparable to RGB stereovision in daylight. Table (d) shows the comparison of practical HADAR ranging (based on TeX-Net outputs) against the optimal HADAR ranging (based on ground truth TeX vision). Practical HADAR ranging shows a near-optimal ranging performance. Table (b) is based on the Street-Long-Animation, Suburb, and Rocky Terrain scenes in the HADAR ranging dataset in the HADAR database. Table (d) is based on the Suburb and Rocky Terrain scenes, as Street-Long-Animation has less spectral bands and is not included in training TeX-Net. TeX-Net statistics were done with 5-fold cross validation.</p> <p>Ground: bottom half of the image. Density: fraction of the overall image area for which statistics is analyzed. Mean error: the mean absolute per-pixel disparity error with respect to the ground truth. Accuracy: fraction of pixels for which the estimated disparity is within τ pixels of the ground truth values.</p>
C8.1	<p>In fig s24, the claim about human vs. robot identification compared to thermal, as seen from the supplementary material, is due to the advantages provided by multi-spectra. As shown in Fig S24, the ability to perform semantic segmentation exploiting the material signature to mask the underlying different material layers followed by detection in that masked layer raises concern about the simultaneous detection of different objects corresponding to different spectra in a scene and performing end-to-end semantic segmentation. Given this strategy, one can directly utilize the multispectral tensor data for this task without TeX-vision, where each spectra can be utilized separately to identify the subject composed of given material (skin vs. aluminum).</p>
R8.1	<p>We would like to clarify before we revise to address this comment.</p> <ol style="list-style-type: none"> The information contained in TeX vision has two main sources. One is the spectro-spatial information from the sensor data (heat cube) mentioned by the reviewer, and <u>the other is the prior information from the material library and the physics model</u>. <ul style="list-style-type: none"> In general, a heat cube contains a mixture of all physical features from temperature, emissivity, and texture. <u>The material library and the physics model help solve the inverse problem and sort the features in an organized representation of TeX vision.</u> <u>Without the material library or the physics model, the issue of TeX degeneracy arises: i.e. a spectrum in the heat cube can be assigned to any material at any temperature.</u> Since mathematically there is no unique solution, the mapping from a heat cube to semantic labels is not well defined. Therefore, an end-to-end segmentation without the physics model will not yield comparable results to TeX vision. We agree that for simple examples/tasks it may be possible to directly use the heat cubes without the physics model for clustering. However, heat cubes are in general mixtures of all features. For example, different temperatures (with the same emissivity and texture, say, robot) will give very different spectra and are likely to lead to different clusters. <u>Clustering based on the total heat cube cannot lead to correct semantic segmentation.</u> <u>TeX vision has the potential to support advanced logical constraints since all the variables are physics-driven.</u> For example, objects cannot be above the sky, temperature of human bodies will be in the vicinity of 37C, trees are not likely to appear on a car even if the reflection gives rise to similar spectra etc.. Those common-sense logic constraints can be utilized on TeX vision for advanced semantic segmentation. However, heat cubes cannot

	<p>support these logic-based constraints on physical variables underlying the task of thermal perception.</p> <p>In summary, one cannot directly use the heat cube without the physics model. Though, we agree that it is not essential to explicitly output TeX vision.</p>
C8.2	<p>To address this concern, can TEX vision enable the simultaneous detection of multi-objects from a single frame without masking?</p>
R8.2	<p>We appreciate the reviewer’s idea about simultaneous detection. We believe that simultaneous detection can be achieved in either of the following approaches. (1) TeX vision images can be used as input to train a neural network for simultaneous detection. In our current work, we used pre-trained networks, and that is why we demonstrated detection sequentially with material regions (masks). (2) Our TeX-Net with the physics model can be utilized as a backbone to design and train novel end-to-end networks for simultaneous detection, and it is not necessary to explicitly output TeX vision. These ideas certainly deserve future studies.</p> <p>In this revised version, we have added the above analysis in Sec.SIIIE of the Supple. Info. as cited below.</p> <p style="text-align: center;">distinguish human vs. robot, which is otherwise impossible. Here, Fig. S24 demonstrates sequential detection by performing detection on each individual material region. We believe that simultaneous detection can be achieved in the following approaches. (1) TeX vision images can be used as input to train a neural network for simultaneous detection. (2) Our TeX-Net with the physics model can be utilized as a backbone to design and train novel</p> <p style="text-align: center;">58</p> <hr style="width: 20%; margin: auto;"/> <p style="text-align: center;">end-to-end networks for simultaneous detection, and it is not necessary to explicitly output TeX vision. These approaches deserve future studies.</p>
C9	<p>On page 71 of supplementary materials, estimating the signatures on the fly for the categorization of materials sounds like a good idea. As the Authors utilize k means to estimate the categories, how do they choose the k as this impacts the categorization process leading to over/underfitting noise during categorization? An incorrect categorization of materials can lead to errors in semantic representations, which propagate to successive tasks. Also, what is a customized TES algorithm?</p>
R9	<p>We would like to address this comment in the following two aspects.</p> <ol style="list-style-type: none"> 1. In this work, our semantic library was estimated by manually choosing a K parameter, as this is our first attempt/step to demonstrate real-world HADAR performance. We agree that this impacts the categorization process and will eventually lead to some errors in the material and semantic maps. This error is, however, inevitable for real-world scenes with open environments where the number of clusters is unknown or difficult to define. As we have observed in our prototype-2 experiments, once the semantic category was set to include tree and grass, bushes (which are in-between these two categories) were predicted as tree or grass. To minimize this error, a potential approach is to scan different K parameters, estimate

semantic library and TeX vision for each K, and then choose the solution with lowest physics loss.

2. Our customized TES algorithm is a modified version of the existing temperature-emissivity separation algorithm that can be found in [1]. The original TES algorithm has a workflow as cited below.

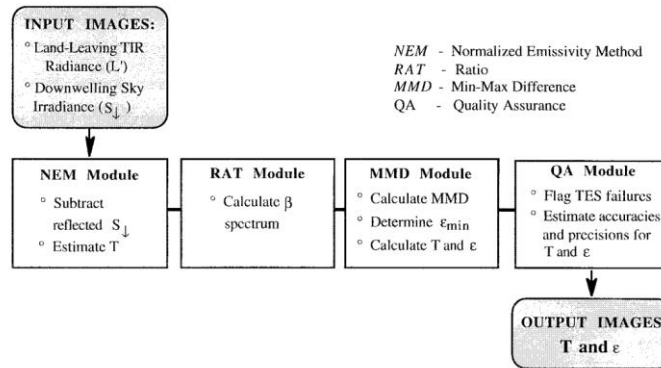


Fig. 1. Basic design of the TES algorithm.

(this figure is only cited for the reviewer and not used in the paper)

The NEM module and RAT module are relevant and adopted in our algorithm. They give the relative profile of the spectral emissivity, leaving one parameter --- the absolute magnitude -- unfixd. After that, the original TES algorithm uses an empirical formula (their Eqs. 5 and 6) to determine temperature and the magnitude of spectral emissivity. Their empirical formula is based on big data from space/air-based applications and hence not applicable to our HADAR experiment related to ground based autonomous navigation.



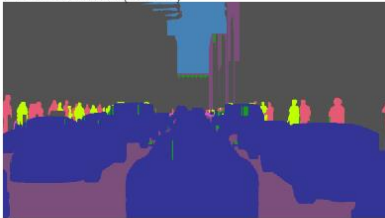

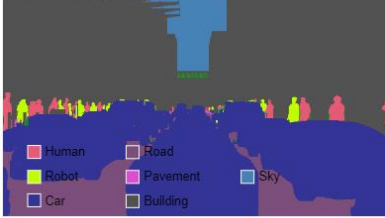
Instead, we use the K-means clustering to categorize materials and derive the averaged emissivity profile for each cluster. We note that multiple pixels of the same cluster (after manual correction to remove unwanted pixels in a cluster) share the same spectral emissivity magnitude. Therefore, we estimate the emissivity magnitude and temperature by least-squares fitting according to our physics model. That completes our customized TES algorithm. Our customized TES algorithm outputs the semantic library, as well as an estimation of the temperature map which is used as an initial solution in the TeX-SGD algorithm.

In this revised version, we have added the above explanation/analysis in Sec.SVC of the Supple. Info. as cited below.

	<p>designed TES (temperature emissivity separation) algorithm to estimate emissivity per pixel. We adopted the NEM and RAT modules from the original TES algorithm that can be found in Ref. [23]. These modules output the relative profile of the spectral emissivity, leaving one parameter – the absolute magnitude – unfixed. After that, the original TES algorithm uses an empirical formula to determine temperature and the magnitude of spectral emissivity. The empirical formula is based on big data from space/air-based applications and hence not applicable to our current HADAR experiments. Instead, we then used the K-means clustering to categorize materials and derive the averaged emissivity profile for each cluster. We note that multiple pixels of the same cluster share the same spectral emissivity magnitude. Therefore, we estimated the emissivity magnitude and temperature by least-squares fitting according to the HADAR constitutional equation. The averaged spectral emissivity for each cluster form the semantic library of the scene. The resulting semantic library is available along with the HADAR database. In this work, we manually chose the K parameter for K-means clustering. This impacts the categorization process and will eventually lead to some errors in the material and semantic maps. To minimize this error, a potential approach deserving future investigations is to scan different K parameters, estimate semantic library and TeX vision for each K, and then choose the solution with lowest physics loss.</p> <p>Reference(s): [1] Gillespie, Alan, et al. "A temperature and emissivity separation algorithm for Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) images." IEEE transactions on geoscience and remote sensing 36.4 (1998): 1113-1126.</p>
C10	<p>Since the authors utilized the result of Tex-SGD as ground truth labels to train the Tex-net, the DL trained is fundamentally limited by the performance of Tex-SGD. Consequently, it is not a fair comparison of the performance of TexNET and TexSGD when the performance of one constrains the performance of the other. The authors should develop alternate loss functions to decouple the dependency or augment some additional form of label for training the Texnet in a supervised fashion.</p>
R10	<p>We agree with the reviewer. We do not intend to compare TeX-Net and TeX-SGD. In this work, we propose TeX-SGD as a non-machine-learning baseline approach, and we treat both TeX-SGD and TeX-Net as novel results to demonstrate HADAR, without the intention to claim any performance ranking between them. We agree that a better approach to train TeX-Net for real-world scenes may be to generate more training data and generate experimental ground truth semantics. In the future, this can be done by manual annotation tools like LabelMe, as other existing visible frequency datasets have done.</p> <p>Since the US army night vision lacks large ground truth experimental data and we do not intend to claim any performance ranking between TeX-SGD and TeX-Net in this paper, <u>we have acknowledged this in the caption of Extended Data Fig.3:</u></p> <p><u>‘...Note that the current TeX-Net was trained partially with TeX-SGD outputs. The above observations are not used to claim performance ranking between TeX-SGD and TeX-Net. Both</u></p>

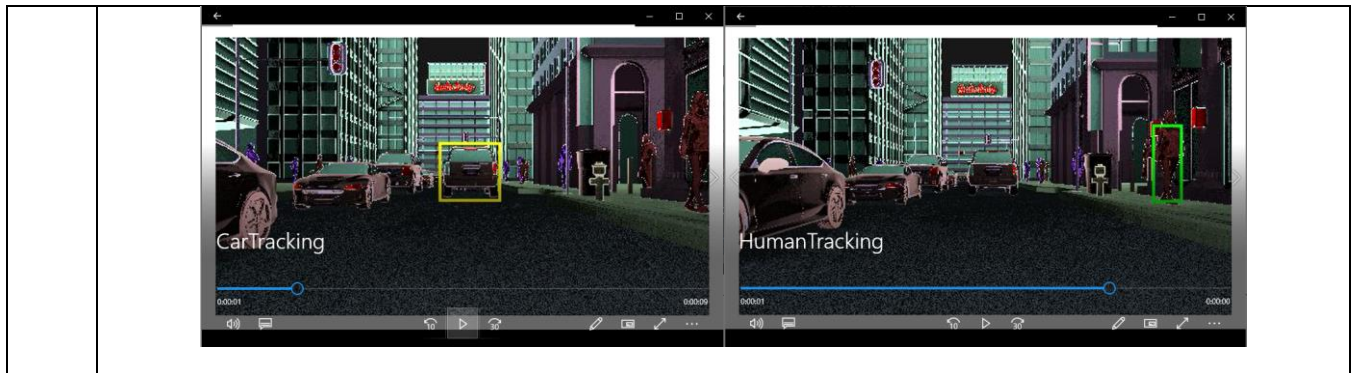
TeX-Net and TeX-SGD confirm that HADAR TeX vision has achieved a semantic understanding of the night scene with enhanced textures comparable to RGB vision in daylight.'

Reviewer 3

C0	<p>The authors have conducted extensive experiments to demonstrate the effectiveness and efficiency of the proposed method. However, I have some minor comments and suggestions that I hope the authors can address before publication. Following are a few suggestions and some questions for the authors:</p>																								
R0	<p>We thank the reviewer for the useful suggestions, as well as all the efforts and patience throughout the whole review process. We have addressed each comment individually below and made corresponding revisions to improve the quality of this manuscript.</p>																								
C1	<p>Most of the scenes in the demo examples provided are relatively easy to segment. I am very curious about the performance in real-world scenes where objects are overlapped or occluded.</p>																								
R1	<p>We regret that our Extended Data Fig.8 about semantic segmentation statistics was not presented in a clear way. In fact, the statistics in the upper table were done on the first 4 on-road scenes in the HADAR database, including Scene 2 -- Crowded Street and Scene 3 -- Suburb (even though Fig.8a-d are only for Scene 1 – Street). <u>Scene 2 and Scene 3 were designed with significant overlap and occlusion, as shown below.</u></p> <div style="text-align: center;">  </div> <p><u>In this latest version, we have revised Extended Data Fig.8a-d to Scene 2 as cited below, to reflect the complexity in our considered scenes in the statistics.</u></p> <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>a Thermal vision + AI (state-of-the-art)</p>  </div> <div style="text-align: center;"> <p>b HADAR semantics (this work)</p>  </div> </div> <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>c Ground truth HADAR material map e(m)</p>  </div> <div style="text-align: center;"> <p>d Ground truth semantics</p>  </div> </div> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <thead> <tr> <th>mIoU %</th> <th>Human</th> <th>Robot</th> <th>Car</th> <th>Road</th> <th>Pavement</th> <th>Building</th> <th>Sky</th> </tr> </thead> <tbody> <tr> <td>Thermal vision + AI</td> <td>33</td> <td>0</td> <td>90</td> <td>67</td> <td>25</td> <td>69</td> <td>16</td> </tr> <tr> <td>TeX vision + non-AI</td> <td>94</td> <td>84</td> <td>98</td> <td>92</td> <td>87</td> <td>84</td> <td>90</td> </tr> </tbody> </table>	mIoU %	Human	Robot	Car	Road	Pavement	Building	Sky	Thermal vision + AI	33	0	90	67	25	69	16	TeX vision + non-AI	94	84	98	92	87	84	90
mIoU %	Human	Robot	Car	Road	Pavement	Building	Sky																		
Thermal vision + AI	33	0	90	67	25	69	16																		
TeX vision + non-AI	94	84	98	92	87	84	90																		

C2	<p>The authors provide a video about the real-world experiment. But the synthesis results are not stable as seen in Tex.avi, for example the trees in the upper right corner of the video keep flashing.</p>
R2	<p>We agree with the reviewer that TeX vision results for the HADAR prototype-2 experiments are not perfect. Here, we would like to explain the relevant details once again before we revise to address this comment.</p> <p>The HADAR prototype-2 real-world experimental data was collected in a collaboration with DARPA (The Defense Advanced Research Projects Agency, through the Invisible Headlights project) and the Army night-vision team (Infrared Camera Technology Branch, DEVCOM C5ISR Center, U.S. Army). This is because the high end sensors which give rise to 256 spectral bands in the thermal infrared spectral range are very expensive and the data collect has to be done in accordance with Army rules for expensive equipment. Please note all the data is available for the broad global audience and there are no restrictions for use by industry or academia.</p> <p>There are multiple practical challenges in experiments, such as, (1) the pushbroom sensor shows horizontal streak noise due to dynamic drift of pixel gain and offset, (2) ground truth material library was not collected, and (3) the sky, which is a significant environmental object, was not directly observed. We have added one section (Sec. SV) in the Supple. Info. to explain the details of denoising, LiDAR-HADAR extrinsic calibration, and estimating the material library as well as the sky radiance. Since the pushbroom sensor was used along with multiple other sensors (irrelevant to this work) in the DARPA IH project, the data collection took so long that we observed significant changes of the estimated sky radiance throughout the experiment. <u>The inaccurate sky radiance estimation causes performance fluctuations of TeX vision, as pointed out by the reviewer. However, we emphasize that all these practical restrictions can be relieved in the future with a proper on-site experimental characterization of the sky radiance and the material library.</u> These limitations are unrelated to the HADAR algorithms itself and we have added additional details as explained in the response to the next question.</p> <p><u>In this revised version, we have added the above analysis in Sec.SVC of the Supple. Info. to acknowledge this limitation, as cited below.</u></p> <p style="text-align: center;"> <small>The sky radiance was not collected in IH experiments as well. We read the heat signal off the reflecting checkerboard (which was facing the sky) to approximate the sky radiance. Since the pushbroom sensor was used along with multiple other sensors (irrelevant to this work) in the IH project, the data collection took so long that we observed significant changes of the estimated sky radiance throughout the experiment. The inaccurate sky radiance estimation causes performance fluctuations of TeX vision. We emphasize that this practical restriction can be relieved with a proper on-site experimental characterization of the sky radiance.</small> </p>

C3	<p>The video of the real-world experiment provided by the authors is short in length and lacks more challenges such as the movement of the objects. Authors are encouraged to provide more video visualizations containing more diversity of scenes.</p>
R3	<p>We agree with the reviewer. As explained in Reply R2, since the real-world experiment is extremely time-consuming and expensive, <u>in this revised version we have provided one more TeX vision video for the synthetic scene --- Street-Long-Animation.</u> The new TeX vision video has 100 frames, much longer than the previous experimental video, and has moving cars. The two TeX vision videos we provided now cover both real-world and numeric experiments, and both on-road and off-road scenes. The videos have also been uploaded to the HADAR database, in the folder ‘Real-world and numeric TeX vision video demonstrations at night’. See below for a screenshot of the new video.</p> <div data-bbox="548 722 1162 1073" data-label="Image"> </div>
C4	<p>Considering that the authors use the ability of object detection to evaluate the quality of TeX vision, the authors should further introduce evaluation of visual object tracking to demonstrate the robustness of TeX vision.</p>
R4	<p>We thank the reviewer for this useful suggestion. In this revised version, we have added the visual object tracking results (a car and a pedestrian) in Sec.SIIIE of the Supple. Info., as cited below.</p> <p style="text-align: center;"> <small>Visual object tracking based on TeX vision and corresponding semantic segmentations has been tested for a car and a pedestrian on the Street-Long-Animation scene. py-tracking implementation (https://github.com/visionml/pytracking) of the ECO [21] method was used in the test. Robust tracking results show the applicability of TeX vision for visual object tracking. Tracking videos are available along with TeX videos at https://github.com/FanglinBao/HADAR.</small> </p>



With these revisions, we believe the manuscript is ready for publication. We thank the reviewers for their input.

Reviewer Reports on the Third Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The reviewer again commends the authors' diligence in addressing the concerns raised in the previous round. The manuscript has significantly improved compared to the first submission, with more analysis, results, and sound and logical discussions. The paper would attract researchers around the world to explore HADAR and exploit this framework for various tasks under low visibility conditions. For all the above reasons, the reviewer would like to recommend the article for publication.

Referee #3 (Remarks to the Author):

The new version of this paper has adequately addressed the issues that have been previously raised by the reviewers. Hence, I recommend it for publication.